Decentralized Non-Convex Optimization and Learning over Heterogeneous Networks

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Ran Xin

Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213

August 2022

Acknowledgements

I want to express my earnest gratitude to my PhD advisor, Soummya Kar, for his wisdom, trust, patience, guidance, and encouragement throughout the years. He brought me to CMU, was always on my side and advocating our work. I greatly appreciate the independence and freedom he provided me in research, which made my life in CMU enjoyable and exciting. I am indebted to my advisor at Tufts, Usman A. Khan. He introduced me to the academic world at the beginning of my graduate study and accompanied me to the end of my PhD journey. I benefited tremendously from numerous discussion with him, beyond research, and it is not an overstatement to say that he shaped me to a better person. Soummya and Usman are the best advisors ever – thank you for everything.

I would like to extend my gratitude to my PhD thesis committee members, Soummya Kar (chair), Usman A. Khan, José M. F. Moura, Anna Scaglione, and Virginia Smith, for their time and invaluable comments in shaping the content of this thesis and my research presentations. I acknowledge my research group members, Carmel Fiscko, Shuhua Yu, Meiyi Li, and Aleksandar Armacki for their insightful feedback in lab meetings and practicing job talks.

I am very grateful to my friends in graduate study. Many thanks to my "basement" fellows, Shuhua Yu, Tian Tong, Jiacheng Zhu, Yingsi Qin, Yuting Deng, Yuhang Yao, Yuwei Qin, Zhuoyuan Wang, Laixi Shi, Hanjiang Hu, Boyue Li, Xiang Wang, and Qin Wang for all the fun and hilarious moments we enjoyed together. I sincerely appreciate Tian Tong, Chenguang Xi, Mengyue Hang, Jianyu Wang, Xiaofei Guo, and Yixuan Yuan, who offered me their generous help in my job search. Special thanks go to Weitong Ruan and Xin Zheng, for their extensive help and constant encouragement along the way. Thank you all – I would not have made it this far without you.

Finally, I thank my parents for their unconditional love and for being my role models. To them I dedicate this thesis.

This thesis is supported partially by NSF under award #1513936 and the Carnegie Institute of Technology Deans Fellowship.

Abstract

We study decentralized optimization and learning problems, where a network of n nodes, such as machines, edge devices, and robot swarms, cooperatively minimizes a finite sum of cost functions by means of local information processing and communication with neighboring nodes. Decentralized optimization has emerged as a promising framework for large-scale machine learning and signal processing problems. It is fundamentally important in scenarios where data samples are geographically distributed and/or centralized data processing is infeasible due to computation and communication overhead or data privacy concerns. Although decentralized optimization has been extensively researched under convexity over the past decade, the field still lacks a sound understanding of how to achieve optimal complexities when the underlying problems of interest become non-convex. In this thesis, we construct provably efficient decentralized stochastic first-order gradient methods for several important classes of non-convex problems with online or offline data, with the help of gradient tracking and variance reduction techniques. In particular, we prove that the proposed algorithms, in regimes of practical significance, achieve network topology-independent computation complexities that match the centralized lower bounds for the corresponding problem classes. This network topology-independence property further leads to the linear speedup of decentralized stochastic optimization algorithms under arbitrary network topologies, in that, the total number of gradient computations at each node is reduced by a factor of $1/n$ compared to the centralized optimal algorithms that perform all gradient computations at a single node. We also discuss several techniques to balance the computation-communication trade-offs in the proposed algorithms. Our algorithmic frameworks and their companion analyses are constructed and developed in a systematic manner and may be generalized to other problems of interest. Extensive numerical experiments with both real and synthetic datasets are included to demonstrate our main theoretical results.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and background

Minimizing a cost function to select an optimal action or decision has been an important problem in science, engineering, and mathematics. The cost function, say $F : \mathbb{R}^p \to \mathbb{R}$, typically quantifies the loss in fitting data or measurements under a model parameterized by $\mathbf{x} \in \mathbb{R}^p$. An optimal model or decision $\mathbf{x}^*$ is often chosen as the one that minimizes the corresponding loss F. Optimization theory and algorithms [5–11] provide the fundamental tools to address such problems. Examples include the classical signal estimation and optimal control problems, where the goal in the former is to minimize the estimation error and in the latter is to minimize the cost of control actions. More recently, with the advent of modern computational machinery, complex nonlinear problems, such as image classification and natural language processing, have enabled a resurgence of interest in the domain of optimization theory and methods.

In this thesis, we investigate decentralized stochastic optimization, where data samples and noisy signal observations are available across multiple nodes, such as machines, sensors, robots, or mobile devices. The nodes communicate with each other according to a peer-to-peer network, without a central coordinator, and solve the underlying optimization problem in a cooperative manner. Such problems are prevalent in modern-day machine learning and signal processing where, for example, a large collection of images are stored on multiple machines in a data center for the purpose of image classification. Moreover, classical applications like sensor networks and robotic swarms also fit this paradigm where the sensors and robots collect measurements in a decentralized manner in order to learn an underlying phenomenon, navigate an environment, or decide on an optimal control action. In such settings, the data samples available at the i-th node lead to a local cost f_i, and the goal of the networked nodes is to *agree* on a minimizer of the global cost $F = \frac{1}{n}\sum_{i=1}^{n} f_i$ based on the data across all n nodes. In related applications of practical interest, raw data

sharing among the nodes is often not permitted due to the private nature of data, such as text messages and medical images, or is inefficient due to limited communication resources. Decentralized first-order optimization methods thus rely on information exchange among the nodes and local gradient computation to build the solution of the global optimization problem. In these methods, each node i retains a local state variable $\mathbf{x}_t^i$ that is an estimate of a minimizer $\mathbf{x}^*$ of the global cost function F at iteration t, and recursively updates this estimate according to the *estimates of the neighboring nodes* and *local gradients*, and perhaps a few other auxiliary variables. In other words, the nodes do not share their raw data directly.

The formulation of decentralized stochastic optimization in general can be divided into two types: (i) *online/streaming data*, where an imprecise (stochastic) gradient is computed based on data samples drawn randomly from an underlying probability distribution at each node; or (ii) *offline/batch data*, where a finite collection of data samples is available locally at each node and a stochastic gradient is computed from samples drawn randomly from the local batch. In this thesis, we develop algorithmic frameworks and complexity results for both online and offline formulations.

1.2 Canonical formulations of decentralized optimization and examples

1.2.1 Canonical formulations

We introduce the canonical forms of the decentralized optimization problems in the following. We consider n nodes, such as machines or edge devices, communicating over a decentralized network described by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \cdots, n\}$ is the set of node indices and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the collection of ordered pairs (i, r), $i, r \in \mathcal{V}$, such that node r sends information to node i. Each node i possesses a private local cost function $f_i : \mathbb{R}^p \to \mathbb{R}$ and the goal of the networked nodes is to solve, via local computation and communication, the following optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}). \tag{1.1}$$

Here, each f_i is only locally accessible and processed by node i and is not shared with any other nodes, since it encodes the local data. We emphasize that the cooperation (information exchange) among the nodes is peer-to-peer without the existence of a central coordinator; see Fig. 1.1. It can be observed that the paradigm of decentralized optimization described above preserves the privacy of local data and achieves data parallelism, thus enabling effective means for flexible parallel computation. In this thesis, we mainly focus on the settings where each local cost function f_i is non-convex, motivated by the applications in deep neural networks [12] and robust learning [13]. Our goal is to design and analyze efficient decentralized optimization algorithms that find a first-order stationary point $\mathbf{x}^*$ of the global cost function F such that $\|\nabla F(\mathbf{x}^*)\| = 0$.

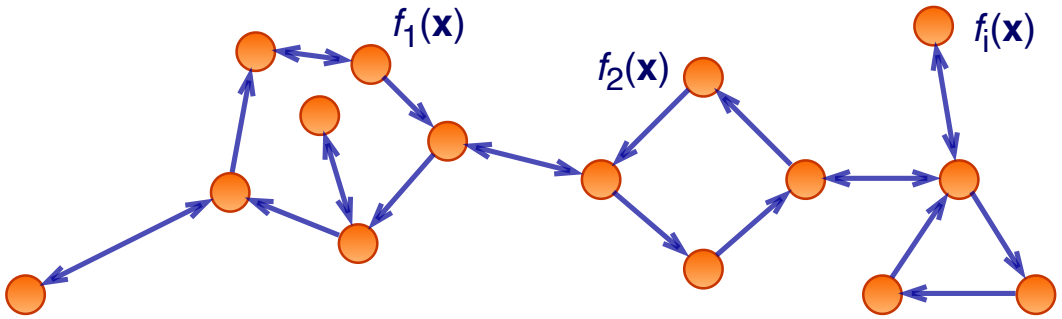

Figure 1.1: Decentralized Optimization: distributed optimization over a graph.

1.2.2 Examples

Problems (1.1) is quite prevalent in signal processing and machine learning problems [14–18]. We provide some representative examples below.

1.2.2.1 Signal-plus-noise model

In classical signal processing, we are often interested in finding an unknown signal $\mathbf{x}^* \in \mathbb{R}^p$ based on the measurements $y_i = \mathbf{h}_i^\top \mathbf{x}^* + v_i$, obtained by a collection of sensors indexed by i, where $\mathbf{h}_i \in \mathbb{R}^p$ is the sensing vector at sensor i and $v_i \in \mathbb{R}$ is the measurement noise. Finding $\mathbf{x}^* \in \mathbb{R}^p$ at each sensor i may be formulated as a local minimization problem in terms of the squared error, i.e., $\min_{\mathbf{x}\in\mathbb{R}^p}(y_i - \mathbf{h}_i^\top \mathbf{x})^2$. However, since this problem may be ill-conditioned and the collected measurements have noise, collaboration among the sensors often leads to a much more robust estimate. The resulting formulation, in the form of Problem (1.1), is

$$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \quad f_i(\mathbf{x}) := (y_i - \mathbf{h}_i^\top \mathbf{x})^2,$$

which is also known as the least-squares problem [19].

1.2.2.2 Expected risk minimization

Problem (1.1) also appears in *(online) expected risk minimization* [7]. In this context, the goal is to find some model $\mathcal{H}$, parameterized by $\mathbf{x} \in \mathbb{R}^p$, that maps an input $\mathbf{z} \in \mathbb{R}^{d_z}$ to its corresponding output $\mathbf{y} \in \mathbb{R}^{d_y}$. The setup requires defining a loss function $\mathcal{L}(\mathcal{H}(\mathbf{z};\mathbf{x}),\mathbf{y})$ that quantifies the mismatch between the model prediction $\mathcal{H}(\mathbf{z};\mathbf{x})$, under the parameter $\mathbf{x}$, and the actual output data $\mathbf{y}$. Assuming that each node i in the network obtains samples in real time from an underlying data stream with distribution $\mathcal{D}_i$, the goal of the networked nodes here is to find the optimal parameter $\mathbf{x}^*$ that minimizes the average of the expected losses across the network, i.e.,

$$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \quad f_i(\mathbf{x}) := \mathbb{E}_{(\mathbf{z}_i,\mathbf{y}_i)\sim\mathcal{D}_i}\left[\mathcal{L}\left(\mathcal{H}(\mathbf{z}_i;\mathbf{x}),\mathbf{y}_i\right)\right]. \tag{1.2}$$

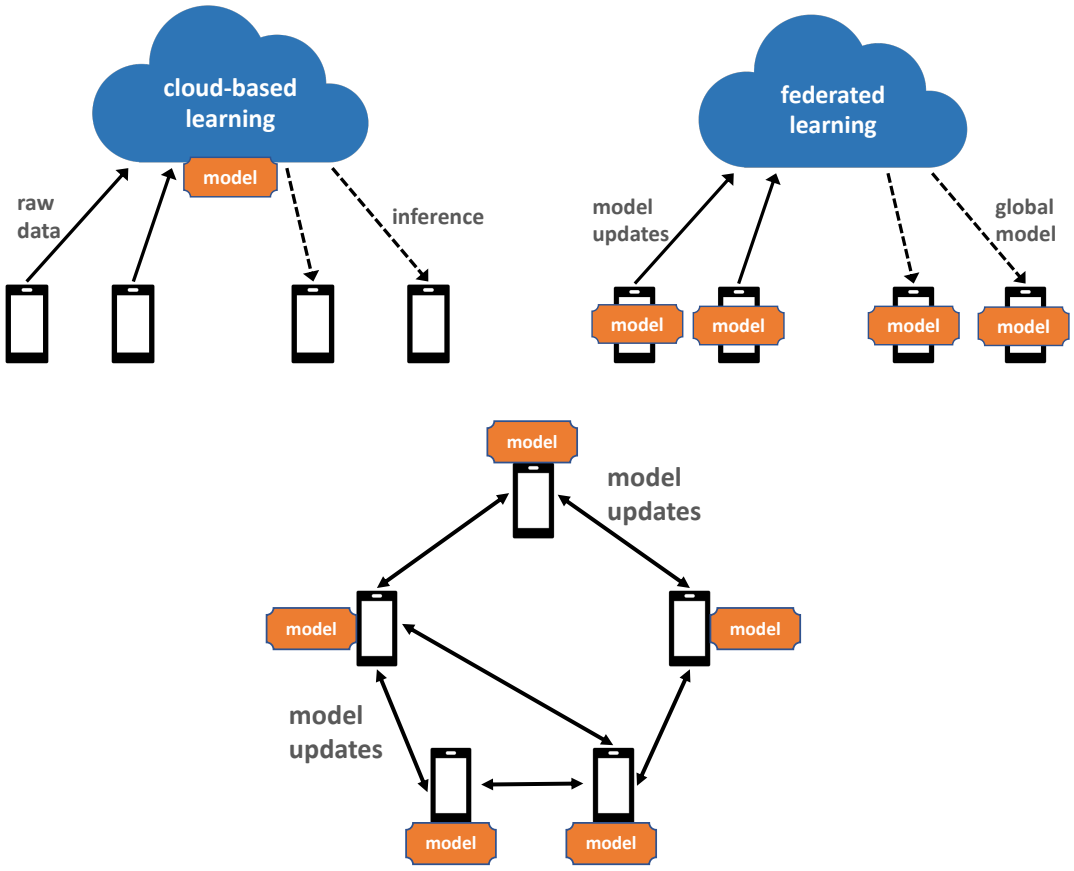

Figure 1.2: An illustration of cloud-based, federated, and decentralized learning in the context of distributed training of machine learning models.

The formulation (1.2) is also known as decentralized online stochastic optimization [4,19–21]. We will provide efficient algorithms and optimal complexity results for this formulation in Chapter 4, 5, and 6.

1.2.2.3 Empirical risk minimization

In practice, each node often has access to a large set of offline data samples $\{(\mathbf{z}_{i,j}, \mathbf{y}_{i,j})\}_{j=1}^{m_i}$ drawn from $\mathcal{D}_i$ described in (1.2), in stead of sampling in real time. In this case, the average loss incurred by all offline data samples across all nodes serves as an appropriate surrogate for the expected risk (1.2) and the resulting problem is often referred as *(offline) empirical risk minimization*, i.e.,

$$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} f_{i,j}(\mathbf{x}), \quad f_{i,j}(\mathbf{x}) := \mathcal{L}\left(\mathcal{H}(\mathbf{z}_{i,j};\mathbf{x}), \mathbf{y}_{i,j}\right). \tag{1.3}$$

The formulation (1.3) is also known as decentralized finite-sum optimization [22–24]. Chapter 2, 3, and 6 discuss how to leverage variance reduction methods to solve this formulation efficiently, where we provide optimal complexity results for non-convex problems.

1.2.3 Advantages over centralized server/worker architectures

In the context of large-scale distributed training of machine learning models, various centralized server/worker programming models and system architectures have been proposed, such as cloud-based learning, MapReduce [25], and federated learning [16], which are tailored for specific computing needs and environments; see Fig. 1.2 for a simple illustration. Such server/worker architectures, although provide scalable solutions, may not be desirable in certain scenarios, described in the following.

- **Communication bottleneck.** When training large-scale machine learning models with a server/worker architecture, the server is often required to constantly push and pull information of very high dimensions from all local workers [26]. In this case, the server could become a communication bottleneck, for instance, in federated learning applications where a massive amount of edge devices cooperatively trains a model of large size over a wireless network [16]. Conversely, the communication in a decentralized network is generally much sparser in the sense that each node only talks to its several neighboring nodes specified by the topology [27]. Decentralized topologies thus achieve faster wall-clock time than the centralized server/worker architectures for each round of communication in network [14, 28, 29]. Leveraging this fact, in this thesis, we demonstrate by rigorous mathematical arguments that decentralized optimization, when properly designed, achieves gradient and communication *complexities*[1] that are comparable to the centralized optimal ones in practical regimes of interest, thus reducing total run time required for training. This outperformance of the decentralized over the centralized is particularly significant if the communication network is of high latency and/or low bandwidth [2].

- **System robustness.** The operation of a server/worker architecture relies heavily on the functionality of the central server. Therefore, the underlying training system may be vulnerable to malicious attacks, as the server is a single point of failure [30]. On the contrary, decentralized optimization methods are applicable as long as the communication network remains connected and are therefore more robust [15].

- **Flexibility.** When enormous data is generated in a local and streaming fashion from a large number of mobile, geographically dispersed, heterogeneous devices, e.g., in the Internet of Things (IoT), one needs a paradigm shift from a server/worker to a peer-to-peer network, since the latter eliminates the need for specialized central coordinators and is based on flexible, non-deterministic, local communication [31].

- **Power consumption.** In federated optimization scenarios, workers typically communicate with the server through long-distance wireless transmission, where the overall power consumption can be quite

[1] The complexities here refer to the total number of gradient computation and communication required for decentralized algorithms to achieve certain accuracy for the underlying optimization problem.

large since the transmission power is often proportional to the squared distance. On the other hand, decentralized networks enable short-distance communication between the nodes, when a geometric nearest-neighbor graph is deployed.[2] As a consequence, decentralized optimization methods are applicable and more efficient in scenarios where the power budget at each node is limited [32].

Motivated these facts, we study decentralized optimization throughout the thesis.

1.3 Decentralized gradient descent and its stochastic variants

In this section, we present and discuss the construction, intuition, and performance of decentralized gradient descent (DGD) and its stochastic variant (DSGD) [19, 21, 33, 34]. It is worth noting that they are the very first decentralized optimization methods and are the prototype of many sophisticated approaches in the field.

1.3.1 The basic average consensus algorithm

At each node i, given the current estimate $\mathbf{x}_t^i$ of the solution at iteration t, related decentralized first-order optimization algorithms typically involve the following steps:

1. compute local (stochastic) gradients of the local function f_i;
2. fuse information with the available neighbors;
3. update the local estimate $\mathbf{x}_{t+1}^i$ according to a specific optimization protocol.

Recall that each node in the network only communicates with its neighbors and only has partial knowledge of the global objective, see Fig. 1.1. Due to this limitation, an information propagation mechanism is required that disseminates local information over the entire network. Decentralized optimization thus has two key components: (i) *agreement or consensus* – all nodes must agree on the same state; and, (ii) *optimality* – the agreement should be on a stationary point of the global objective F. Average-consensus algorithms [35] are information fusion protocols that enable each node to appropriately combine the vectors received from its neighbors and to agree on the average of the initial states of the nodes. They thus naturally serve as basic building blocks in decentralized optimization, added to which are local gradient corrections that steer the agreement to a global stationary point.

To describe average-consensus, we first associate the communication graph with a primitive and doubly-stochastic weighted adjacency matrix $\underline{\mathbf{W}} = \{\underline{w}_{ir}\} \in \mathbb{R}^{n\times n}$, such that $\underline{w}_{ir} \neq 0$ if and only if node r sends information to node i in the graph. Clearly, we have $\mathbf{W}\mathbf{1}_n = \mathbf{1}_n$ and $\mathbf{W}^\top \mathbf{1}_n = \mathbf{1}_n$, where $\mathbf{1}_n \in \mathbb{R}^n$ is the

[2] In geometric graphs, two nodes are connected if and only if they are in physical vicinity.

column vector of n ones. There are various ways of constructing such weights in a decentralized manner. Popular choices include the Laplacian and Metropolis weights; see, e.g., [27] for details. The basic average-consensus algorithm [35] is given as follows. For all $t \geq 0$, each node i starts with some vector $\mathbf{x}_0^i \in \mathbb{R}^p$ and updates its state according to

$$\mathbf{x}_{t+1}^i = \sum_{r=1}^{n} w_{ir} \mathbf{x}_t^r.$$

This update can be written in a vector form as

$$\mathbf{x}_{t+1} = (\underline{\mathbf{W}} \otimes \mathbf{I}_p)\mathbf{x}_t, \tag{1.4}$$

where $\mathbf{x}_t = [\mathbf{x}_t^{1\top}, \cdots, \mathbf{x}_t^{n\top}]^\top \in \mathbb{R}^{np}$ and $\otimes$ denotes the Kronecker product. Since $\underline{\mathbf{W}}$ is primitive and doubly-stochastic, from the Perron-Frobenius theorem [36], we have

$$\lim_{t\to\infty} \underline{\mathbf{W}}^t = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$$

and therefore the states in average consensus follow

$$\lim_{t\to\infty} \mathbf{x}_t = \lim_{t\to\infty} (\underline{\mathbf{W}} \otimes \mathbf{I}_p)^t \mathbf{x}_0 = \mathbf{1}_n \otimes \overline{\mathbf{x}}_0$$

where

$$\overline{\mathbf{x}}_0 := \frac{(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{x}_0}{n}.$$

In other words, the average consensus protocol in (1.4) enables agreement among the nodes on the average $\overline{\mathbf{x}}_0$ of the initial states. It can be further shown that (1.4) converges at a linear rate of λ^k [36], where $\lambda \in [0, 1)$ is the second largest singular value of $\underline{\mathbf{W}}$.

1.3.2 Construction of DSGD and its basic intuitions

With the agreement protocol (1.4) in place, we now introduce the well-known decentralized gradient descent (DGD) built on top of it and provide basic intuitions. The update rule of DGD [21, 33, 37] is described as follows. Each node i starts with an arbitrary $\mathbf{x}_0^i \in \mathbb{R}^p$ and updates, for all $t \geq 0$, according to

$$\mathbf{x}_{t+1}^i = \sum_{r=i}^{n} \underline{w}_{ir} \mathbf{x}_t^r - \alpha_t \nabla f_i\left(\mathbf{x}_t^i\right), \tag{1.5}$$

where α_t is the step-size. Indeed, DGD adds local gradient corrections to average-consensus (1.4). In order to understand the iterations of DGD, we write them in a vector form. Let $\mathbf{x}_t$ and $\nabla\mathbf{f}(\mathbf{x}_t)$ collect all local states and gradients, respectively, i.e., $\mathbf{x}_t := [\mathbf{x}_t^{1\top}, \cdots, \mathbf{x}_t^{n\top}]^\top$ and $\nabla\mathbf{f}(\mathbf{x}_t) := [\nabla f_1(\mathbf{x}_t^1)^\top, \cdots, \nabla f_n(\mathbf{x}_t^n)^\top]^\top$, both in $\mathbb{R}^{np}$. Then the update of DGD can be compactly written as

$$\mathbf{x}_{t+1} = (\underline{\mathbf{W}} \otimes \mathbf{I}_p)\mathbf{x}_t - \alpha_t \nabla\mathbf{f}(\mathbf{x}_t). \tag{1.6}$$

We further define the average $\overline{\mathbf{x}}_t := \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{x}_t$ of the local states at time t and multiply both sides of (1.6) by $\frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)$ to obtain:

$$\overline{\mathbf{x}}_{t+1} = \overline{\mathbf{x}}_t - \alpha_t \frac{(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\nabla \mathbf{f}\left(\mathbf{x}_t\right)}{n}. \tag{1.7}$$

Based on (1.6) and (1.7), it can be observed that the consensus matrix $\underline{\mathbf{W}}$ makes the states $\{\mathbf{x}_t^i\}_{i=1}^n$ at the nodes approach their average $\overline{\mathbf{x}}_t$, while the average gradient $\frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\nabla \mathbf{f}\left(\mathbf{x}_t\right)$ steers $\overline{\mathbf{x}}_t$ towards a stationary point of F. The overall DGD protocol thus ensures agreement and optimality simultaneously, the two key components of decentralized optimization as we described before.

For large-scale machine learning and signal processing problems, DSGD, a stochastic variant of DGD, is often used in practice [19, 37–39]. The basic update of DSGD is described as follows. Let $\mathbf{x}_t^i \in \mathbb{R}^p$ denote the state at node i and iteration t. For all $t \geq 0$, DSGD performs

$$\mathbf{x}_{t+1}^i = \sum_{r=1}^n \underline{w}_{ir}\mathbf{x}_t^r - \alpha_t \mathbf{g}_t^i, \tag{1.8}$$

where $\mathbf{g}_t^i \in \mathbb{R}^p$ is a stochastic gradient such that $\mathbb{E}[\mathbf{g}_t^i|\mathbf{x}_t^i] = \nabla f_i(\mathbf{x}_t^i)$. DSGD is popular for several inference and learning tasks due to its simplicity of implementation and speedup in comparison to centralized SGD algorithms [2]. DSGD and its variants have been extensively studied for different computation and communication requirements, e.g., momentum [40], directed graphs [41], escaping saddle-points [42, 43], zeroth-order schemes [44], swarming-based implementations [45], and constrained problems [46].

1.3.3 Challenges with DSGD

When each f_i is non-convex, the performance of DSGD however suffers from two major challenges:

- the non-degenerate variance of the stochastic gradients at each node;
- the heterogeneity among the local functions/data across the nodes.

To elaborate these issues, we recap the convergence results of DSGD (1.8) when each local function f_i is smooth and non-convex. Let us assume the bounded variance of each local stochastic gradient $\mathbf{g}_t^i$, the *bounded heterogeneity* between the local and the global gradient [2], i.e., for some $\nu > 0$ and $\zeta > 0$,

$$\sup_{i,t} \mathbb{E}\left[\left\|\mathbf{g}_t^i - \nabla f_i(\mathbf{x}_t^i)\right\|^2\right] \leq \nu^2 \qquad \text{and} \qquad \sup_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}) - \nabla F(\mathbf{x})\right\|^2 \leq \zeta^2,$$

and the L-smoothness of each f_i. Then it is shown in [2] that, DSGD achieves an ϵ-accurate stationary point $\mathbf{x}^*$ of the global function F such that $\mathbb{E}[\|\nabla F(\mathbf{x}^*)\|] \leq \epsilon$ in

$$O\left(\frac{L\nu^2}{n\epsilon^4} + \frac{nL^2\nu^2}{(1-\lambda)\epsilon^2} + \frac{nL^2\zeta^2}{(1-\lambda)^2\epsilon^2}\right) \tag{1.9}$$

iterations, where $\lambda \in [0,1)$ is the second largest singular value of the network weight matrix $\underline{\mathbf{W}}$. The complexity bound of DSGD in (1.9) is the summation of three terms: the first term matches the iteration complexity of the centralized SGD with minibatch size n [47]; the second term reveals how the decentralized network topology affects the run time of DSGD; and the last term reveals the adversarial impact of heterogeneous data on DSGD. Clearly, there are two major issues with the convergence properties of DSGD:

- The bounded heterogeneity assumption on the local and global gradients [2, 41, 43] or the coercivity of each local function [42] is essential for establishing the convergence of DSGD. In fact, a counterexample has been shown in [15] that *DSGD diverges for any constant step-size* when these types of assumptions are violated. Furthermore, the theoretical and practical performance of DSGD degrades significantly when the local and the global gradients are substantially different, i.e., when the data distributions across the nodes are largely heterogeneous [3, 4, 20].
- Due to the non-degenerate stochastic gradient variance, the gradient complexity of DSGD does not match the centralized lower bounds for several fundamental classes of stochastic and finite-sum non-convex optimization problems [48–50].

This thesis designs and analyzes new decentralized optimization algorithms that improve the complexity bounds and practical performance of DSGD by tackling the above issues. In particular, we address the first issue by the gradient tracking technique, e.g., [51–56], described in the next section, and the second issue with the help of variance reduction schemes, e.g., [48–50, 57–64].

1.4 The gradient tracking technique

The gradient tracking technique was proposed to address the impact of heterogeneous data across the nodes in the convergence of decentralized optimization methods [51–56, 65] and is a key ingredient of the algorithms proposed in this thesis. To present the intuition behind the gradient tracking technique, we first recall the iterations of the (non-stochastic) Decentralized Gradient Descent (DGD) with a constant step-size in (1.5). Let us first assume, for the sake of argument, that all nodes agree on a stationary point $\mathbf{x}^*$ of the global function F at some iteration t, i.e., $\mathbf{x}_t^i = \mathbf{x}^*, \forall i$. Then at the next iteration $t+1$, we have

$$\mathbf{x}_{t+1}^i = \sum_{r=1}^{n} w_{ir}\mathbf{x}^* - \alpha \nabla f_i(\mathbf{x}^*) = \mathbf{x}^* - \alpha \nabla f_i(\mathbf{x}^*) \neq \mathbf{x}^*, \tag{1.10}$$

where the equality uses the fact that $\underline{\mathbf{W}} = \{\underline{w}_{ir}\}$ is doubly-stochastic and the last equality holds because $f_i \neq F$ in general. In other words, $\mathbf{x}^*$ is not necessarily a fixed point of the DGD algorithm. Of course, replacing the local gradient $\nabla f_i\left(\mathbf{x}_t^i\right)$ in the DGD algorithm with the gradient $\nabla F\left(\mathbf{x}_t^i\right)$ of the *global* function

overcomes this issue but the global gradient is not available at any node due to the decentralized topology. The natural yet innovative idea of gradient tracking is to design a local iterative gradient tracker $\mathbf{y}_t^i$ that asymptotically approaches the global gradient $\nabla F\left(\mathbf{x}_t^i\right)$ as $\mathbf{x}_t^i$ approaches $\mathbf{x}^*$ [52–56, 65]. Gradient tracking is implemented with the help of dynamic average consensus (DAC) [51], briefly described next.

In contrast to classical average-consensus [35] that converges to the average of fixed initial states, DAC [51] tracks the average of time-varying signals. Formally, each node i measures a time-varying signal $\mathbf{s}_t^i$ and the goal of all nodes is to collaboratively track the average $\overline{\mathbf{s}}_t := \frac{1}{n}\sum_{i=1}^n \mathbf{s}_t^i$ of these signals. For all $t \geq 0$, the DAC protocol is given as follows. Each node i iteratively updates its estimate $\mathbf{y}_t^i$ of $\overline{\mathbf{s}}_t$ as

$$\mathbf{y}_{t+1}^i = \sum_{r=1}^n w_{ir}\mathbf{y}_t^r + \mathbf{s}_{t+1}^i - \mathbf{s}_t^i, \tag{1.11}$$

where $\mathbf{y}_0^i = \mathbf{s}_0^i$ for all i. It is shown in [51] that if $\lim_{t\to\infty}\left\|\mathbf{s}_{t+1}^i - \mathbf{s}_t^i\right\| = 0$, then we have that

$$\lim_{t\to\infty}\left\|\mathbf{y}_t^i - \overline{\mathbf{s}}_t\right\| = 0.$$

Clearly, in the aforementioned design of gradient tracking, the time-varying signal that we intend to track is the average of the local gradients $\frac{1}{n}\sum_{i=1}^n \nabla f_i\left(\mathbf{x}_t^i\right)$. We thus combine DGD (1.5) and DAC (1.11) to obtain *GT-DGD (DGD with Gradient Tracking)* [52, 54–56, 65], as follows:

$$\mathbf{x}_{t+1}^i = \sum_{r=1}^n w_{ir}\mathbf{x}_t^r - \alpha\cdot\mathbf{y}_t^i, \tag{1.12a}$$

$$\mathbf{y}_{t+1}^i = \sum_{r=1}^n w_{ir}\mathbf{y}_t^r + \nabla f_i\left(\mathbf{x}_{t+1}^i\right) - \nabla f_i\left(\mathbf{x}_t^i\right), \tag{1.12b}$$

where $\mathbf{y}_0^i = \nabla f_i\left(\mathbf{x}_0^i\right)$ for all i. Intuitively, as $\mathbf{x}_t^i \to \overline{\mathbf{x}}_t$ and $\mathbf{y}_t^i \to \frac{1}{n}\sum_{i=1}^n \nabla f_i\left(\mathbf{x}_t^i\right) \to \nabla F\left(\overline{\mathbf{x}}_t\right)$, (1.12a) thus asymptotically becomes the centralized batch gradient descent. GT-DGD consistently outperforms DGD in heterogeneous data settings [14]. Although GT-DGD has been widely researched under convexity, see, e.g., [36, 52–54, 66, 67], its performance for the stochastic non-convex settings remains unclear. We address this gap in Chapter 4. It is worth noting that all the algorithms proposed in this thesis use the gradient tracking technique, in order to address the issue of heterogeneous local cost functions.

1.5 A brief literature review

Decentralized optimization, also known as distributed optimization over graphs, starts with several well-known papers on decentralized gradient descent (DGD) [21, 33, 34, 37, 38], where the focus was primarily on signal processing and control problems defined in sensor and robotic networks. Since then many decentralized methods with more sophisticated algorithmic structures have been proposed to improve the performance of DGD from various computation and communication aspects. Decentralized first-order gradient methods that

improve the performance of DGD include, e.g., EXTRA [68], Exact Diffusion/NIDS [3, 20, 69], DLM [70], and methods based on gradient tracking [4, 54–56, 65, 67, 71, 72]; see also primal-dual frameworks [66, 73–75] that unify the aforementioned methods under certain conditions. Dual gradient methods can be found in [71, 76, 77]. Decentralized second-order algorithms that leverage curvature information of the cost functions to accelerate the convergence can be found in [78–82]. Alternating direction method of multipliers (ADMM) and its variants have also been used in decentralized optimization [83–89]. References [90–97] design algorithms that adapt to communication and computation imperfection and trade-offs, e.g., time-varying and random graphs, asynchronous execution, and quantization. In the context of decentralized non-convex optimization, there have been results beyond finding first-order stationary points. For instance, [98] develops decentralized annealing methods for finding a global minima in certain regularized non-convex problems. References [42, 43, 99, 100] establish convergence of related algorithms to second-order stationary points. The work [101] studies a family of non-convex non-smooth problems with the help of the generalized gradient.

The last decade has witnessed a vastly growing literature in the area of decentralized stochastic, convex, and non-convex optimization problems; we invite the readers to, e.g., survey articles [14, 15, 27, 31, 102] and the references therein, in addition to the work discussed above. *Despite this fact, it appears that the field still lacks a sound theory and understanding on how to achieve optimal gradient and communication complexities for decentralized non-convex optimization under various stochastic settings.* This thesis addresses this gap.

1.6 Contributions

The overarching theme of this thesis is to provide optimal gradient and communication complexity results for finding *first-order stationary points* in several fundamental classes of decentralized stochastic non-convex problems. In particular, we find various regimes of practical significance where the gradient complexity of the proposed decentralized algorithms matches the centralized lower bound for the corresponding problem classes. While retaining the optimal gradient complexities, we also achieve optimal communication complexities by means of the mini-batch technique and multi-round accelerated consensus algorithms. In the following, we describe the contribution of each chapter in this thesis.

1.6.1 Smooth strongly-convex finite-sum problems (Chapter 2)

In this chapter, we describe a novel algorithmic framework to construct decentralized stochastic variance-reduced methods. The proposed framework, which we call `GT-VR`, is stochastic and decentralized, and thus is particularly suitable for problems where large-scale, potentially private data, cannot be collected or processed at a centralized server. The `GT-VR` framework leads to a family of algorithms with two key ingredients: (i)

local variance reduction, that enables estimation of the local exact gradients from randomly drawn samples of local data with reduced variance; and, (ii) *global gradient tracking*, which fuses the local gradient information across the nodes to track the global gradient. Naturally, the integration of different variance reduction and gradient tracking techniques leads to different algorithms of interest with valuable practical trade-offs and design considerations. For instance, Chapter 2, 3, 5, and 6 respectively apply this framework to different classes of optimization problems of interest.

In the context of smooth strongly convex functions, we focus on two instantiations of the `GT-VR` framework, namely `GT-SAGA` and `GT-SVRG`, that exhibit a compromise between space and time. We show that both `GT-SAGA` and `GT-SVRG` achieve linear convergence to the optimal solution for smooth and strongly convex problems and further describe the regimes in which they achieve non-asymptotic, network topology-independent linear rates that are faster with respect to the existing decentralized first-order schemes. Moreover, we show that both algorithms achieve a linear speedup in such regimes, in that, the total number of gradient computations required at each node is reduced by a factor of $1/n$, where n is the number of nodes, compared to their centralized counterparts that process all data at a single node.

1.6.2 Smooth non-convex finite-sum problems (Chapter 3)

In this chapter, we consider decentralized minimization of $N := nm$ smooth non-convex cost functions equally divided over a network of n nodes, where each node possesses a local batch of m cost functions, i.e., data samples. In this context, we propose two decentralized stochastic variance-reduced gradient methods, under the `GT-VR` framework described in Section 1.6.1, that achieve provably fast and robust convergence.

1.6.2.1 Stochastic recursive variance reduction

We propose `GT-SARAH` that employs a `SARAH`-type variance reduction technique and gradient tracking (`GT`) to address the stochastic and decentralized nature of the problem. We show that `GT-SARAH`, with appropriate algorithmic parameters, finds an ϵ-stationary point with $\mathcal{O}\big(\max\big\{N^{1/2}, n(1-\lambda)^{-2}, n^{2/3}m^{1/3}(1-\lambda)^{-1}\big\}L\epsilon^{-2}\big)$ gradient complexity, where $(1-\lambda) \in (0,1]$ is the spectral gap of the network weight matrix and L is the smoothness parameter of the cost functions. This gradient complexity outperforms that of the existing decentralized stochastic gradient methods. In particular, in a big-data regime such that $n = \mathcal{O}(N^{1/2}(1-\lambda)^3)$, this gradient complexity furthers reduces to $\mathcal{O}(N^{1/2}L\epsilon^{-2})$, independent of the network topology, and matches that of the centralized optimal variance-reduced methods. Moreover, in this regime `GT-SARAH` achieves a *non-asymptotic linear speedup*, in that, the total number of gradient computations at each node is reduced by a factor of $1/n$ compared to the centralized optimal algorithms that perform all gradient computations

at a single node. To the best of our knowledge, GT-SARAH is the first algorithm that achieves this property. In addition, we show that appropriate choices of local minibatch size balance the trade-offs between the gradient and communication complexity of GT-SARAH. Over infinite time horizon, we establish that all nodes in GT-SARAH asymptotically achieve consensus and converge to a first-order stationary point in the almost sure and mean-squared sense.

1.6.2.2 Stochastic incremental variance reduction

We analyze the performance of the GT-SAGA algorithm proposed in Chapter 2 in the non-convex settings. For general smooth non-convex problems, we show the almost sure and mean-squared convergence of GT-SAGA to a first-order stationary point of the global cost function, and further describe regimes of practical significance where it outperforms the existing approaches and achieves a network topology-independent iteration complexity respectively. When the global function satisfies the Polyak-Lojaciewisz condition, we show that GT-SAGA exhibits linear convergence to an optimal solution in expectation and describe regimes of practical interest where the performance is network topology-independent and improves upon the existing methods.

1.6.3 Smooth non-convex online stochastic problems (Chapter 4)

In this chapter, we study decentralized non-convex optimization, where each node accesses its local function by means of an online stochastic first-order oracle. Integrating the gradient tracking technique in decentralized stochastic gradient descent, we show that the resulting algorithm, GT-DSGD, enjoys certain desirable characteristics towards minimizing a sum of smooth non-convex costs. In particular, for general smooth non-convex functions, we establish non-asymptotic characterizations of GT-DSGD and derive the conditions under which it achieves network topology-independent performances that match the centralized minibatch SGD. In contrast, the existing results suggest that GT-DSGD is always network topology-dependent and is therefore strictly worse than the centralized minibatch SGD. When the global function additionally satisfies the Polyak-Lojasiewics (PL) condition, we establish the linear rate of GT-DSGD up to a steady-state error with appropriate constant step-sizes. Moreover, under stochastic approximation step-sizes, we establish, for the first time, the optimal global sublinear convergence rate on almost every sample path, in addition to the asymptotically optimal sublinear rate in expectation. Since strongly convex functions are a special case of the functions satisfying the PL condition, our results are not only immediately applicable but also improve the currently known best convergence rates and their dependence on problem parameters.

1.6.4 Non-convex online stochastic problems with mean-squared smoothness (Chapter 5)

In this chapter, we study decentralized non-convex optimization, where each node accesses its local function by means of an online stochastic first-order oracle that satisfies a mean-squared smoothness property. In this context, we propose, under the `GT-VR` framework described in Section 1.6.1, a novel single-loop decentralized hybrid variance-reduced stochastic gradient method, called `GT-HSGD`, that outperforms the existing approaches in terms of both the gradient complexity and practical implementation. `GT-HSGD` implements specialized local hybrid stochastic gradient estimators that are fused over the network to track the global gradient. Remarkably, `GT-HSGD` achieves a network topology-independent oracle complexity of $O(n^{-1}\epsilon^{-3})$ when the required error tolerance ϵ is small enough, leading to a linear speedup with respect to the centralized optimal approaches for this problem class that operate on a single node.

1.6.5 Non-convex non-smooth composite problems (Chapter 6)

In this chapter, we focus on decentralized non-convex composite problems over networked nodes, where the network cost is the average of local smooth non-convex risks plus an extended valued, convex, possibly non-differentiable regularizer. To the best of our knowledge, the existing decentralized stochastic optimization literature lacks non-asymptotic gradient and communication complexity results for this composite problem formulation. In this chapter, we address this gap by introducing a unified framework, called `ProxGT`, that is built upon local stochastic gradient estimators and a global gradient tracking technique. We construct several different instantiations of this framework by choosing appropriate local estimators for the corresponding problem classes. In particular, we develop `ProxGT-SA` and `ProxGT-SR-O` for the expected risk, and `ProxGT-SR-E` for the empirical risk. Remarkably, we show that each algorithm achieves a network topology-independent optimal gradient complexity with an optimal communication complexity for the corresponding problem class.

1.7 Practical concerns and future directions

The major scope of this thesis is to achieve optimal gradient and communication complexities for decentralized optimization in several classical and fundamental classes of stochastic non-convex problems. With the help of these complexity results, we provide a theoretical justification to the fact that decentralized methods can outperform the corresponding centralized ones for various machine learning and signal processing tasks. In the following, we discuss some limitations of the convergence theory developed in this thesis from the view of practical applications and implementations.

- **Deep learning models.** The complexity theory developed in this thesis heavily relies on the Lipschitz smoothness assumption [6] made for the local cost functions that encode the underlying model structure and local data, while we allow the existence of extended-valued non-smooth convex regularization such as ℓ_1-norm and/or general closed convex constraints. This smoothness assumption holds, e.g., for the family of non-convex generalized linear models with convex regularizers [13]. However, for complex nonlinear problems such as large-scale natural language processing and image classification tasks, deep neural networks like LSTM [103], ResNets [104], and Transformers [105] are often deployed in practice. These deep learning models typically use highly convoluted layers with the non-smooth activation functions and therefore their associated cost functions do not satisfy the Lipschitz smoothness assumption in general. Despite this fact, recent studies have shown empirical success of decentralized methods in training state-of-the-art deep learning models [2,28,29]. Therefore, it may be beneficial and interesting to establish convergence theory of decentralized optimization methods for deep models by looking into their specific structures, where classical optimization theory may not apply directly [106].

- **Model generalization.** Throughout this thesis, we treat the problem of training machine learning models from pure optimization perspectives, i.e., we establish gradient and communication complexities to find ϵ-accurate stationary points [7] of the global cost function. That is to say, we primarily use the gradient norm as the convergence metric. On the other hand, test accuracy (generalization capability) is typically used to evaluate the quality of a machine learning model in practice and hence may be a more informative metric [13]. As a future direction, it is interesting to establish statistical generalization bounds [13] and perform large-scale experiments for decentralized optimization methods and further examine whether the intuitions developed in this thesis, such as network topology-independent performance, hold true in the sense of model generalization and test accuracy.

- **Communication imperfections.** In the proposed decentralized stochastic optimization algorithms and their companion complexity theory, we do not take communication imperfection into account, such as asynchronous execution of the nodes [107], stragglers [108], time-varying and random topolgies [91, 109], channel noise [110], model and gradient compression techniques [111], quantization [21], and package losses [112]. However, in modern-day applications like training large models on heterogeneous mobile devices and the Internet of Things (IoT), handling the aforementioned issues is essential from a practical implementation standard point, since they significantly affect the run time of the underlying optimization methods and the quality of the resulting solution. It is hence advantageous to adapt the algorithms and convergence results established in this thesis to these more practical settings.

Chapter 2

Decentralized Smooth Strongly-Convex Finite-Sum Optimization

This chapter describes a novel algorithmic framework, **GT-VR**, for decentralized smooth stochastic problems. Specifically, **GT-VR** subsumes a family of efficient algorithms with two major components: (i) *variance reduction*, which produces variance-reduced local exact gradient estimates from random samples of local data; and, (ii) *gradient tracking*, which fuses local gradient information across the nodes. It is clear that using different variance reduction and gradient tracking techniques in **GT-VR** leads to different constructions of decentralized optimization methods with valuable trade-offs of practical interest.

In this chapter, we demonstrate the performance of the **GT-VR** framework by focusing on its two instantiations, which we call **GT-SAGA** and **GT-SVRG**, for smooth and strongly-convex problems.[1] It is shown that both **GT-SAGA** and **GT-SVRG** achieve accelerated linear convergence to the optimal solution for this problem class. We further identify the regimes where they achieve non-asymptotic, network topology-independent linear rates that are faster with respect to the existing decentralized gradient schemes. Moreover, we show that both algorithms achieve a linear speedup in such regimes, in that, the total number of gradient computations required at each node is reduced by a factor of $1/n$, where n is the number of nodes, compared to the centralized SAGA and SVRG algorithms that process all data at a single node. Extensive simulations are presented to illustrate the performance of the proposed algorithms.

[1] In the later chapters, we will discuss other instantiations of **GT-VR** tailored for various non-convex problems.

2.1 Introduction

In this chapter, we consider decentralized finite-sum minimization problems that take the following form:

$$\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \quad f_i(\mathbf{x}) := \frac{1}{m_i}\sum_{j=1}^{m_i} f_{i,j}(\mathbf{x}), \tag{2.1}$$

where each cost function $f_i : \mathbb{R}^p \to \mathbb{R}$ is private to a node i, in a network of n nodes, and is further subdivided into an average of m_i component functions $\{f_{i,j}\}_{j=1}^{m_i}$. This formulation has found tremendous interest over the past decade and has been studied extensively by the signal processing, control, and machine learning communities [2, 33, 34]. When the dataset is large-scale and further contains private information, it is often not feasible to communicate and process the entire dataset at a central location. Decentralized stochastic gradient methods thus are preferable as they not only benefit from local (short-range) communication but also exhibit low computation complexity by sampling and processing small subsets of data at each node i, instead of the entire local batch of m_i functions.

Decentralized stochastic gradient descent (DSGD) was introduced in [19, 21, 39], which combines network fusion with local stochastic gradients and has been popular in various decentralized learning tasks. However, the performance of DSGD is mainly adversely impacted by two components: (i) the variance of the local stochastic gradients at each node; and, (ii) the dissimilarity between the datasets and local functions across the nodes. In this chapter, we propose a novel algorithmic framework, namely **GT-VR**, that systematically addresses both of these aspects of DSGD by building an estimate of the global descend direction $-\nabla F$ locally at each node based on local stochastic gradients. In particular, the **GT-VR** framework leads to a family of algorithms with two key ingredients: (i) *local variance reduction*, that estimates the local batch gradients $\sum_j \nabla f_{i,j}$ from arbitrarily drawn samples of local data; and, (ii) *global gradient tracking*, which uses the aforementioned local batch gradient estimates and fuses them across the nodes to track the global batch gradient $\sum_i \nabla f_i$. Naturally, existing methods for variance reduction, such as SAG [113], SAGA [57], SVRG [61], SARAH [58], and for gradient tracking, such as dynamic average consensus [51, 54–56] and dynamic average diffusion [114], are all valid choices for the two components in **GT-VR** and lead to various design choices and practical trade-offs.

In this chapter, we focus on smooth and strongly convex problems, where simple schemes, such as SAGA and SVRG, are shown to obtain linear convergence and strong performance. These two methods are extensively studied in the centralized settings and exhibit a compromise between space and time. Specifically, SAGA in practice demonstrates faster convergence compared with SVRG [31, 57], however at the expense of additional storage requirements. Consequently, we consider the following two instantiations of the **GT-VR** framework: (i) **GT-SAGA**, which is an incremental gradient method that requires $\mathcal{O}(pm_i)$ storage

cost at each node i; and, (ii) **GT-SVRG**, which is a hybrid gradient method that does not require additional storage but computes local batch gradients periodically, which leads to stringent requirements on network synchronization and may add latency to the overall implementation.

2.1.1 Related work

Significant progress has been made recently towards decentralized first-order gradient methods. Examples include EXTRA [68], Exact-Diffusion [115], methods based on gradient-tracking [52–56, 65] and primal-dual methods [70, 73]; these full gradient methods, based on certain bias-correction principles, achieve linear convergence to the optimal solution for smooth and strongly convex problems and improve upon the well-known DGD [33], where a constant step-size leads to linear but inexact convergence. Several stochastic variants of EXTRA, Exact-Diffusion, and gradient tracking methods have been recently studied in [2–4, 4, 20, 43, 67, 116]; these methods, due to the non-diminishing variance of the local stochastic gradients, converge sub-linearly to the optimal solution with decaying step-sizes and outperform their deterministic counterparts when local data batches are large and low-precision solutions suffice [31]. Exact linear convergence to the optimal solution has been obtained with the help of variance reduction where existing decentralized stochastic methods include [22, 23, 117–120]. The proposed **GT-VR** framework leads to accelerated convergence over the related stochastic methods; a detailed comparison will be conducted with the help of numerical simulations.

2.1.2 Main contributions

We enlist the main contributions of this chapter as follows:

- We describe **GT-VR**, *a novel algorithmic framework* to minimize a finite sum of functions over a decentralized network of nodes.
- Focusing on two particular instantiations of **GT-VR**, **GT-SAGA** and **GT-SVRG**, we show how different combinations of variance reduction and gradient tracking potentially lead to valuable practical considerations in terms of storage, computation, and communication tradeoffs.
- We show that both **GT-SAGA** and **GT-SVRG** achieve accelerated linear convergence to the optimal solution for smooth and strongly convex problems.
- We characterize the regimes in which **GT-SAGA** and **GT-SVRG** achieve non-asymptotic, network-independent convergence rates and exhibit a linear speedup, in that, the total number of gradient computations at each node is reduced by a factor of $1/n$ compared to their centralized counterparts that process all data at a single node.

To the best of our knowledge, **GT-SAGA** and **GT-SVRG** are the first decentralized stochastic methods that show *provable network-independent linear convergence* and *linear speedup* without requiring the expensive computation of dual gradients or proximal mappings of the cost functions.

The rest of this chapter is structured as follows. Section 2.2 develops the class of decentralized stochastic variance-reduced algorithms proposed in this chapter while Section 2.3 presents the main convergence results and a detailed comparison with the state-of-the-art. Section 2.4 provides extensive numerical simulations to illustrate the convergence behavior of the proposed methods. Section 2.5 presents a unified approach to cast and analyze the proposed algorithms, where Sections 2.5.3 and 2.5.4 contain the convergence analysis for **GT-SAGA** and **GT-SVRG**, respectively. Section 2.6 concludes this chapter.

We use lowercase bold letters to denote vectors and $\|\cdot\|$ to denote the Euclidean norm of a vector. The matrix $\mathbf{I}_d$ is the $d \times d$ identity, and $\mathbf{1}_d$ (resp. $\mathbf{0}_d$) is the d-dimensional column vector of all ones (resp. zeros). For two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d\times d}$, $\mathbf{X} \otimes \mathbf{Y}$ denotes their Kronecker product. The spectral radius of a matrix $\mathbf{X}$ is denoted by $\rho(\mathbf{X})$, while its spectral norm is denoted by $|||\mathbf{X}|||$. The weighted infinity norm of $\mathbf{y} = [y_1, \cdots, y_d]^\top$ given a positive weight vector $\mathbf{x} = [x_1, \cdots, x_d]^\top$ is defined as $\|\mathbf{y}\|_\infty^{\mathbf{x}} = \max_i |y_i|/x_i$ and $|||\cdot|||_\infty^{\mathbf{x}}$ is the matrix norm induced by $\|\cdot\|_\infty^{\mathbf{x}}$.

2.2 Development of the GT-VR framework

In this section, we systematically build the proposed **GT-VR** framework and describe its two instantiations, **GT-SAGA** and **GT-SVRG**. To this aim, we consider DSGD [19, 21, 39], a well-know decentralized version of stochastic gradient descent, and its convergence guarantee for smooth and strongly convex problems as follows. Let $\mathbf{x}^*$ denote the unique minimizer of Problem (2.1) and $\mathbf{x}_i^k \in \mathbb{R}^p$ denote the estimate of $\mathbf{x}^*$ at node i and iteration k of DSGD. The update of DSGD is given by

$$\mathbf{x}_i^{k+1} = \sum_{r=1}^{n} \underline{w}_{ir}\mathbf{x}_r^k - \alpha \cdot \nabla f_{i,s_i^k}(\mathbf{x}_i^k), \qquad k \geq 0, \tag{2.2}$$

where the matrix $\underline{\mathbf{W}} = \{\underline{w}_{ir}\} \in \mathbb{R}^{n\times n}$ collects the weights that each node assigns to its neighbors and the index s_i^k is chosen uniformly at random from the set $\{1, \ldots, m_i\}$ at each iteration k. Assuming bounded variance of $\nabla f_{i,s_i^k}(\mathbf{x}_i^k)$, i.e., $\mathbb{E}[\|\nabla f_{i,s_i^k}(\mathbf{x}_i^k) - \nabla f_i(\mathbf{x}_i^k)\|^2 \mid \mathbf{x}_i^k] \leq \nu^2, \forall i, k$, and cost functions to be smooth and strongly convex, it can be shown that with an appropriate constant step-size α the mean-squared error $\mathbb{E}[\|\mathbf{x}_i^k - \mathbf{x}^*\|^2]$, at each node i, decays linearly up to a steady state error such that [20]

$$\limsup_{k\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|^2\right] = \mathcal{O}\left(\frac{\alpha\nu^2}{n\mu} + \frac{\alpha^2 Q^2 \nu^2}{1-\lambda} + \frac{\alpha^2 Q^2 \zeta^2}{(1-\lambda)^2}\right), \tag{2.3}$$

where $\zeta^2 := \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}^*)\|_2^2$, $(1-\lambda)$ is the spectral gap of the network weight matrix $\underline{\mathbf{W}}$, and Q is the condition number of F. This steady-state error, due to the presence of ν^2 and ζ^2, can be eliminated with

the help of decaying step-size $\alpha_k = \mathcal{O}(1/k)$; however, the convergence rate becomes sub-linear $\mathcal{O}(1/k)$ [116]. In other words, there is an inherent rate/accuracy trade-off in the performance of `DSGD`. The proposed **`GT-VR`** framework, built on global gradient tracking and local variance reduction, completely removes the steady-state error of `DSGD` and achieves fast convergence with a constant step-size to the exact solution.

The proposed **`GT-VR`** framework combines two well-known techniques from recent centralized and decentralized optimization literature to systematically eliminate the steady-state error of `DSGD` and as a consequence recovers linear convergence to the exact solution. The framework has two key ingredients:

(i) *Local Variance Reduction:* **`GT-VR`** removes the performance limitation due to the variance ν^2 of the stochastic gradients by asymptotically estimating the local batch gradient ∇f_i, at each node i, based on randomly drawn samples from the local data. Many variance reduction schemes, e.g., [57, 58, 61, 113], are applicable here and a suitable one can be chosen based on the underlying problem specifications.

(ii) *Global Gradient Tracking:* The other error source ζ^2 is due to the fact that $\nabla f_i(\mathbf{x}^*) \neq \mathbf{0}_p, \forall i$, in general, because of the difference between the local and global cost functions. This issue is addressed with the help of gradient tracking techniques [51, 54–56] that properly fuse the local batch gradient estimates (obtained from the local variance reduction procedures described above) to track the global batch gradient.

Our focus in this chapter is on smooth and strongly convex problems for which the variance reduction methods SAGA [57] and SVRG, in centralized settings, are shown to achieve strong practical performance and theoretical guarantees. These two methods contrast each other, in that, they can be viewed as a compromise between space and time [57], where SAGA requires additional storage but, in practice, demonstrates faster convergence as compared to SVRG, where additional storage is not required. Additionally, the two methods are build upon different variance-reduction principles, i.e., SAGA is a randomized incremental gradient method, whereas SVRG is a hybrid gradient method that evaluates batch gradients periodically in addition to stochastic gradient computations at each iteration, as will be detailed further. We thus explicitly focus on these two methods in this chapter, formally described next.

2.2.1 The `GT-SAGA` algorithm

Algorithm 1 formally describes the SAGA-based implementation of **`GT-VR`**. To implement the gradient estimator, each node i maintains a table of component gradients $\{\nabla f_{i,j}(\mathbf{z}_{i,j}^k)\}_{j=1}^{m_i}$, where $\mathbf{z}_{i,j}^k$ is the most recent iterate at which the component gradient $\nabla f_{i,j}$ was evaluated up to iteration k. At each iteration k, each node i samples an index s_i^k uniformly at random from the local indices $\{1, \cdots, m_i\}$ and computes its local

Algorithm 1 `GT-SAGA` at each node i

Require: $\mathbf{x}_i^0$; $\mathbf{z}_{i,j}^1 = \mathbf{x}_i^0,\ \forall j \in \{1, \cdots, m_i\}$; α; $\{\underline{w}_{ir}\}_{r=1}^n$; $\mathbf{y}_i^0 = \mathbf{g}_i^0 = \nabla f_i(\mathbf{x}_i^0)$.

1: **for** $k = 0, 1, 2, \cdots$ **do**

2: Update the local estimate of the solution:

$$\mathbf{x}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{x}_r^k - \alpha \mathbf{y}_i^k;$$

3: Select s_i^{k+1} uniformly at random from $\{1, \cdots, m_i\}$;

4: Update the local gradient estimator:

$$\mathbf{g}_i^{k+1} = \nabla f_{i,s_i^{k+1}}\big(\mathbf{x}_i^{k+1}\big) - \nabla f_{i,s_i^{k+1}}\big(\mathbf{z}_{i,s_i^{k+1}}^{k+1}\big) + \frac{1}{m_i}\sum_{j=1}^{m_i}\nabla f_{i,j}\big(\mathbf{z}_{i,j}^{k+1}\big);$$

5: If $j = s_i^{k+1}$, then $\mathbf{z}_{i,j}^{k+2} = \mathbf{x}_{i,j}^{k+1}$; else $\mathbf{z}_{i,j}^{k+2} = \mathbf{z}_{i,j}^{k+1}$.

6: Update the local gradient tracker:

$$\mathbf{y}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{y}_r^k + \mathbf{g}_i^{k+1} - \mathbf{g}_i^k;$$

7: **end for**

gradient estimator as

$$\mathbf{g}_i^k = \nabla f_{i,s_i^k}\big(\mathbf{x}_i^k\big) - \nabla f_{i,s_i^k}\big(\mathbf{z}_{i,s_i^k}^k\big) + \frac{1}{m_i}\sum_{j=1}^{m_i}\nabla f_{i,j}\big(\mathbf{z}_{i,j}^k\big).$$

After $\mathbf{g}_i^k$ is computed, the s_i^k-th element in the gradient table is replaced by $\nabla f_{i,s_i^k}(\mathbf{x}_i^k)$, while other entries remain unchanged. The local estimators $\mathbf{g}_i^k$'s are then fused over the network to compute $\mathbf{y}_i^k$, which tracks the global batch gradient ∇F at each node i, and is used as the descent direction to update the local estimate $\mathbf{x}_i^k$ of the optimal solution. Clearly, each local estimator $\mathbf{g}_i^k$ approximates the local batch gradient ∇f_i in an incremental manner via the average of the past component gradients in the table. This implementation procedure results in a storage cost of $\mathcal{O}(pm_i)$ at each node i, which can be reduced to $\mathcal{O}(m_i)$ for certain structured problems [57, 113].

2.2.2 The `GT-SVRG` algorithm

Algorithm 2 formally describes the SVRG-based implementation of **`GT-VR`**. In contrast to **`GT-SAGA`** that incrementally approximates the local batch gradients via past component gradients, **`GT-SVRG`** achieves variance reduction by evaluating the local batch gradients ∇f_i's *periodically*. **`GT-SVRG`** may be interpreted as a "double loop" method, where each node i, at every outer loop update $\{\mathbf{x}_i^{tT}\}_{t\geq 0}$, calculates a local full gradient $\nabla f_i(\mathbf{x}_i^{tT})$ that is retained in the subsequent inner loop iterations to update the local gradient

Algorithm 2 `GT-SVRG` at each node i

Require: $\mathbf{x}_i^0$; $\boldsymbol{\tau}_i^0 = \mathbf{x}_i^0$; α; $\{\underline{w}_{ir}\}_{r=1}^n$; T; $\mathbf{y}_i^0 = \mathbf{v}_i^0 = \nabla f_i(\mathbf{x}_i^0)$.

1: **for** $k = 0, 1, 2, \cdots$ **do**

2: Update the local estimate of the solution:

$$\mathbf{x}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{x}_r^k - \alpha \mathbf{y}_i^k;$$

3: Select s_i^{k+1} uniformly at random from $\{1, \cdots, m_i\}$;

4: If $\mathrm{mod}(k+1, T) = 0$, then $\boldsymbol{\tau}_i^{k+1} = \mathbf{x}_i^{k+1}$, else $\boldsymbol{\tau}_i^{k+1} = \boldsymbol{\tau}_i^k$;

5: Update the local stochastic gradient estimator:

$$\mathbf{v}_i^{k+1} = \nabla f_{i,s_i^{k+1}}\big(\mathbf{x}_i^{k+1}\big) - \nabla f_{i,s_i^{k+1}}\big(\boldsymbol{\tau}_i^{k+1}\big) + \nabla f_i\big(\boldsymbol{\tau}_i^{k+1}\big);$$

6: Update the local gradient tracker:

$$\mathbf{y}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{y}_r^k + \mathbf{v}_i^{k+1} - \mathbf{v}_i^k;$$

7: **end for**

estimator $\mathbf{v}_i^k$, i.e., for $k \in [tT, (t+1)T - 1]$,

$$\mathbf{v}_i^k = \nabla f_{i,s_i^k}\big(\mathbf{x}_i^k\big) - \nabla f_{i,s_i^k}\big(\mathbf{x}_i^{tT}\big) + \nabla f_i\big(\mathbf{x}_i^{tT}\big).$$

Clearly, `GT-SVRG` eliminates the requirement of storing the most recent component gradients at each node and thus has a favorable storage cost compared with `GT-SAGA`. However, this advantage comes at the expense of evaluating two stochastic gradients $\nabla f_{i,s_i^k}\big(\mathbf{x}_i^k\big)$ and $\nabla f_{i,s_i^k}\big(\mathbf{x}_i^{tT}\big)$ at every iteration, in addition to calculating the local batch gradients ∇f_i's every T iterations. See Remarks 2.3.1 and 2.3.2 for additional discussion.

2.3 Main convergence results

The convergence results for `GT-SAGA` and `GT-SVRG` are established under the following assumptions.

Assumption 2.3.1. *The global cost function F is μ-strongly convex, i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ and for some $\mu > 0$, we have*

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \big\langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \big\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

We note that under Assumption 1, the global cost function F has a unique minimizer, denoted as $\mathbf{x}^*$.

Assumption 2.3.2. *Each local cost function $f_{i,j}$ is L-smooth, i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ and for some $L > 0$, we have*

$$\|\nabla f_{i,j}(\mathbf{x}) - \nabla f_{i,j}(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

Clearly, under Assumption 2.3.2, the global cost F is also L-smooth and $L \geq \mu$. We use $Q \coloneqq L/\mu$ to denote the condition number of the global cost F.

Assumption 2.3.3. *The weight matrix $\underline{\mathbf{W}} = \{\underline{w}_{ir}\}$ associated with the network $\mathcal{G}$ is primitive and doubly stochastic.*

Assumption 2.3.3 is not only restricted to undirected graphs and is further satisfied by the class of strongly-connected directed graphs that admit doubly stochastic weights. This assumption implies that the second largest singular value λ of $\underline{\mathbf{W}}$ is less than 1, i.e, $\lambda = \left\| \underline{\mathbf{W}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \right\| < 1$ [36]. Note that although we focus on the basic case of static networks which appear, for instance, in data centers, the convergence analysis provided here can be possibly extended to the more general case of time-varying dynamic networks following the methodology in [91].

We denote $M \coloneqq \max_i m_i$ and $m \coloneqq \min_i m_i$, where m_i is the number of local component functions at node i. The main convergence results of **GT-SAGA** and **GT-SVRG** are summarized respectively in the following theorems.

Theorem 2.3.1 (Mean-square convergence of **GT-SAGA**)**.** *Let Assumptions 2.3.1, 2.3.2, and 2.3.3 hold. If the step-size α in **GT-SAGA** is such that*

$$\alpha \asymp \min\left\{\frac{1}{\mu M}, \frac{m(1-\lambda)^2}{MLQ}\right\},$$

then we have: $\forall k \geq 0,\ \forall i \in \{1, \cdots, n\}$,

$$\mathbb{E}\left[\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|^2\right] \lesssim \left(1 - \min\left\{\frac{1}{M}, \frac{m(1-\lambda)^2}{MQ^2}\right\}\right)^k.$$

***GT-SAGA** thus achieves an ϵ-optimal solution of $\mathbf{x}^*$ in*

$$\mathcal{O}\left(\max\left\{M, \frac{MQ^2}{m(1-\lambda)^2}\right\}\log\frac{1}{\epsilon}\right)$$

component gradient computations (iterations) at each node.

Theorem 2.3.2 (Mean-square convergence of **GT-SVRG**)**.** *Let Assumptions 2.3.1, 2.3.2, and 2.3.3 hold. If the step-size α and the length T of the inner loop are such that*

$$\alpha \asymp \frac{(1-\lambda)^2}{LQ}, \qquad T \asymp \frac{Q^2 \log Q}{(1-\lambda)^2},$$

then we have: $\forall t \geq 0,\ \forall i \in \{1, \cdots, n\}$,

$$\mathbb{E}\left[\left\|\mathbf{x}_i^{tT} - \mathbf{x}^*\right\|^2\right] \lesssim 0.7^t$$

GT-SVRG *thus achieves an ϵ-optimal solution of $\mathbf{x}^*$ in*

$$\mathcal{O}\left(\left(M+\frac{Q^2\log Q}{(1-\lambda^2)^2}\right)\log\frac{1}{\epsilon}\right)$$

component gradient computations at each node.

Theorems 2.3.1 and 2.3.2 lead to the following linear convergence rates for **GT-SAGA** and **GT-SVRG** on almost every sample path, following directly from Chebyshev's inequality and the Borel-Cantelli lemma; see Lemma 2.5.7 for details.

Corollary 2.3.1 (Almost sure convergence of **GT-SAGA**). *Let Assumptions 2.3.1, 2.3.2 and 2.3.3 hold. For the choice of the step-size α in Theorem 2.3.1, we have: $\forall i\in\{1,\cdots,n\}$,*

$$\mathbb{P}\left(\lim_{k\to\infty}\gamma_g^{-k}\left\|\mathbf{x}_i^k-\mathbf{x}^*\right\|^2=0\right)=1,$$

where

$$\gamma_g\asymp 1-\min\left\{\frac{1}{M},\frac{m(1-\lambda)^2}{MQ^2}\right\}.$$

Corollary 2.3.2 (Almost sure convergence of **GT-SVRG**). *Let Assumptions 2.3.1, 2.3.2 and 2.3.3 hold. For the choice of the step-size α and the length T of the inner loop in Theorem 2.3.2, we have: $\forall i\in\{1,\cdots,n\}$,*

$$\mathbb{P}\left(\lim_{t\to\infty}(0.7+\delta)^{-t}\left\|\mathbf{x}_i^{tT}-\mathbf{x}^*\right\|^2=0\right)=1,$$

where $\delta>0$ is an arbitrary small constant.

We discuss some salient features of the proposed algorithms next and compare them with the state-of-the-art.

Remark 2.3.1 (**Big data regime**). When each node has a large dataset such that $M\approx m\gg Q^2(1-\lambda)^{-2}$, we note that both **GT-SAGA** and **GT-SVRG**, achieve an ϵ-optimal solution with a network-independent component gradient computation complexity of $\mathcal{O}(M\log\frac{1}{\epsilon})$ at each node; in contrast, centralized SAGA and SVRG, that process all data on a single node, require $\mathcal{O}((nM+Q)\log\frac{1}{\epsilon})\approx\mathcal{O}(nM\log\frac{1}{\epsilon})$ component gradient computations [57,61]. **GT-SAGA** and **GT-SVRG** therefore achieve a non-asymptotic, linear speedup in this big data regime, i.e., the number of component gradient computations required per node is reduced by a factor of $1/n$ compared with their centralized counterparts[2].

Remark 2.3.2 (**GT-SAGA versus GT-SVRG**). It can be observed from Theorems 2.3.1 and 2.3.2 that when data samples are unevenly distributed across the nodes, i.e., $\frac{M}{m}\gg 1$, **GT-SVRG** achieves a lower gradient

[2]We emphasize that linear speedup, although desirable and somewhat plausible, is not necessarily achieved for decentralized methods in general. In other words, the advantage of parallelizing an algorithm over n nodes may not naturally result into a performance improvement of n.

computation complexity than **GT-SAGA**. However, an uneven data distribution may adversely impact the practical implementation of **GT-SVRG**. This is because **GT-SVRG** requires a highly synchronized communication network as all nodes need to evaluate their local batch gradients every T iterations and cannot proceed to the next inner loop until all nodes complete this local computation. As a result, the nodes with smaller datasets have a relatively long idle time at the end of each inner loop that leads to an increase in overall wall-clock time. Indeed, the inherent trade-off between **GT-SAGA** and **GT-SVRG** is the network synchrony versus the gradient storage. For structured problems, where the component gradients can be stored efficiently, **GT-SAGA** may be preferred due to its flexibility of implementation and less dependence on network synchronization. Conversely, if the problem of interest is large-scale, i.e., m is very large, and storing all component gradients is not feasible, **GT-SVRG** may become a more appropriate choice.

Remark 2.3.3 (**Communication complexities**). Note that since **GT-SAGA** incurs $\mathcal{O}(1)$ communication round per node at each iteration, its total communication complexity is the same as its iteration complexity, i.e., $\mathcal{O}\big(\max\big\{M, \frac{M}{m}\frac{Q^2}{(1-\lambda)^2}\big\}\log\frac{1}{\epsilon}\big)$. For **GT-SVRG**, we note that a total number of $\mathcal{O}(\log\frac{1}{\epsilon})$ outer-loop iterations are required, where each outer-loop iteration incurs $T = \mathcal{O}\left(\frac{Q^2\log Q}{(1-\lambda)^2}\right)$ rounds of communication per node, leading to a total communication complexity of $\mathcal{O}\left(\frac{Q^2\log Q}{(1-\lambda)^2}\log\frac{1}{\epsilon}\right)$. Clearly, in a big data regime where each node has a large dataset, **GT-SVRG** achieves a lower communication complexity than **GT-SAGA**.

Remark 2.3.4 (**Comparison with Related Work**). Existing decentralized variance-reduced (VR) gradient methods include: DSA [22] that integrates EXTRA [68] with SAGA [57] and was the first decentralized VR method; DAVRG that combines Exact Diffusion [115] and AVRG [121]; DSBA [117] that uses proximal mapping [122] to accelerate DSA; Ref. [118] that applies edge-based method [123] to DSA; and ADFS [119] that is a decentralized version of the accelerated randomized proximal coordinate gradient method [124] based on the dual of Problem (2.1). Both **GT-SAGA** and **GT-SVRG** improve upon the convergence rates in terms of the joint dependence on Q and m for these methods, especially in the "big data" scenarios where m is very large, with the exception of DSBA and ADFS. We note that DSBA [117] and ADFS [119], both achieve better a gradient computation complexity albeit at the expense of computing the proximal mapping of a component function at each iteration that is in general very expensive. Another recent work [120] considers gradient tracking and variance reduction and proposes a decentralized SVRG type algorithm. However, the convergence of the decentralized SVRG in [120] is only established when the local functions are sufficiently similar. In contrast, **GT-SAGA** and **GT-SVRG** proposed in this chapter achieve accelerated linear convergence for arbitrary local functions and are robust to the heterogeneity of local functions and data distributions. Finally, we emphasize that all existing decentralized VR methods require symmetric weights and thus undirected networks. In contrast, **GT-SAGA** and **GT-SVRG** only require doubly stochastic weights and therefore can

Table 2.1: Summary of datasets used in numerical experiments. All datasets are available in LIBSVM [1].

Dataset	**train** ($N = nm$)	**dimension** (p)	**test**
Fashion-MNIST	10,000	784	4,000
Covertype	400,000	54	181,012
CIFAR-10	10,000	3,072	2,000
Higgs	90,000	28	8,050
a9a	32,560	123	16,282
w8a	49,740	300	14,960

be implemented over directed graphs that admit doubly stochastic weights [125], providing a more flexible topology design.

2.4 Numerical Experiments

In this section, we numerically demonstrate the convergence behavior of **GT-SAGA** and **GT-SVRG** under different regimes of interest and compare their performances with the-state-of-the-art decentralized stochastic first-order algorithms under different graph topologies and datasets. We consider a decentralized training problem where a network of n nodes with m data samples locally at each node cooperatively finds a regularized logistic regression model for binary classification:

$$F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\log\left[1+e^{-(\mathbf{x}^\top \boldsymbol{\theta}_{ij})\xi_{ij}}\right] + \frac{\lambda}{2}\|\mathbf{x}\|_2^2,$$

where $\boldsymbol{\theta}_{ij} \in \mathbb{R}^p$ denotes the feature vector of the j-th data sample at the i-th node, $\xi_{ij} \in \{-1, +1\}$ is the corresponding binary label, and λ is a regularization parameter to prevent overfitting of the training data. The datasets in question are summarized in Table 2.1 and all feature vectors are normalized to be unit vectors, i.e., $\|\boldsymbol{\theta}_{ij}\| = 1, \forall i, j$. The graph topologies under considerations, shown in Fig 2.1, are directed ring graphs, directed exponential graphs, and undirected nearest-neighbor geometric graphs, all with self loops. We note that the directed ring graph has the weakest connectivity among all strongly-connected graphs; directed exponential graphs, where each node sends information to the nodes $2^0, 2^1, 2^2, \cdots$ hops away, are sparse yet well-connected and therefore are often preferable when one has the freedom to design the graph topology; undirected nearest-neighbor geometric graphs, where two nodes are connected if they are in physical vicinity, are weakly-connected and often arise in ad hoc settings such as robotics swarms, IoTs, and edge computing networks. The doubly stochastic weights for directed ring and exponential graphs are chosen as uniform weights, while the weights for geometric graphs are generated by the Metropolis rule [27]. The parameters

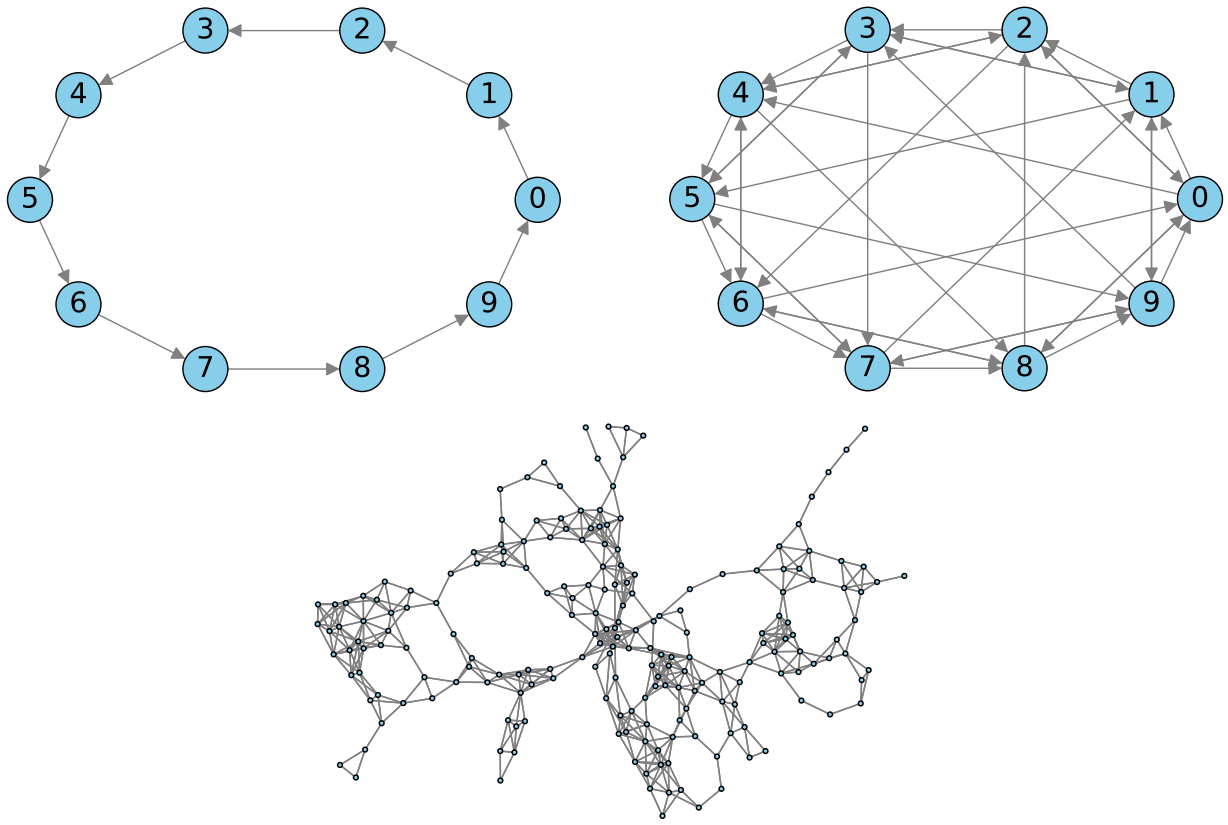

Figure 2.1: The directed ring graph with 10 nodes, directed exponential graph with 10 nodes, and an undirected geometric graph with 200 nodes.

of all algorithms in all cases are manually tuned for best performance. We characterize the performance of the decentralized optimization methods in question in terms of the optimality gap $\frac{1}{n}\sum_{i=1}^{n}\left(F(\mathbf{x}_k^i) - F(\mathbf{x}^*)\right)$ and model accuracy on the test data sets over epochs, where we assume that each node possesses the same number m of data samples and one epoch represents m gradient computations per node.

2.4.1 Big data regime: topology-independence and linear speedup

In this subsection, we demonstrate the convergence behavior of **GT-SAGA** and **GT-SVRG** in the big data regime, i.e., $m \approx Q^2(1-\sigma)^{-2}$. To this aim, we choose 500,000 training samples from the Covertype dataset, equally distributed in a network of $n = 10$ nodes such that each node has $m = 50{,}000$ data samples and set the regularization parameter as $\lambda = 0.01$ that leads to $Q \approx 25$, where Q is the condition number of F. We test the performance of **GT-SAGA** and **GT-SVRG** over different graph topologies, i.e., the directed ring, the directed exponential, and the complete graph with 10 nodes; the second largest singular eigenvalues of the weight matrices associated with these three graphs are $\sigma = 0.951, 0.6, 0$, respectively. It can be verified that the big data condition holds for the optimization problem defined on these three graphs. The experimental results are shown in Fig. 2.2 (left and middle) and we observe that, in this big data regime, the convergence rates of

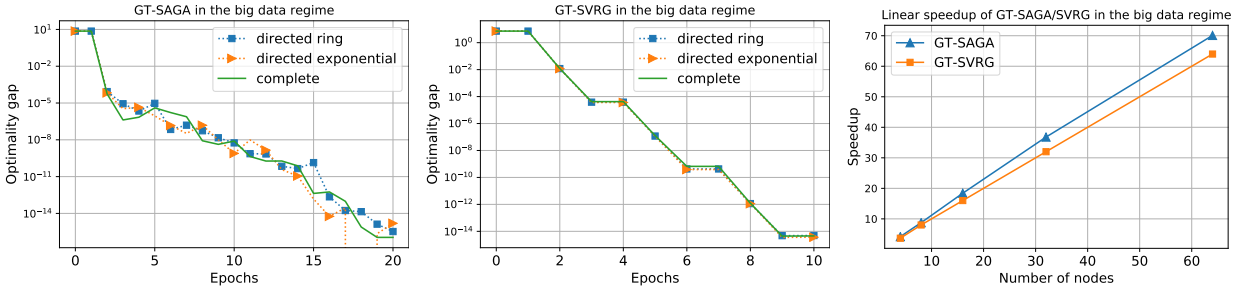

Figure 2.2: The convergence behavior of **GT-SAGA** and **GT-SVRG** in the big data regime: (Left and Middle) Non-asymptotic, network-independent convergence; (Right) Linear speedup with respect to centralized **SAGA** and **SVRG** that process all data on a single node.

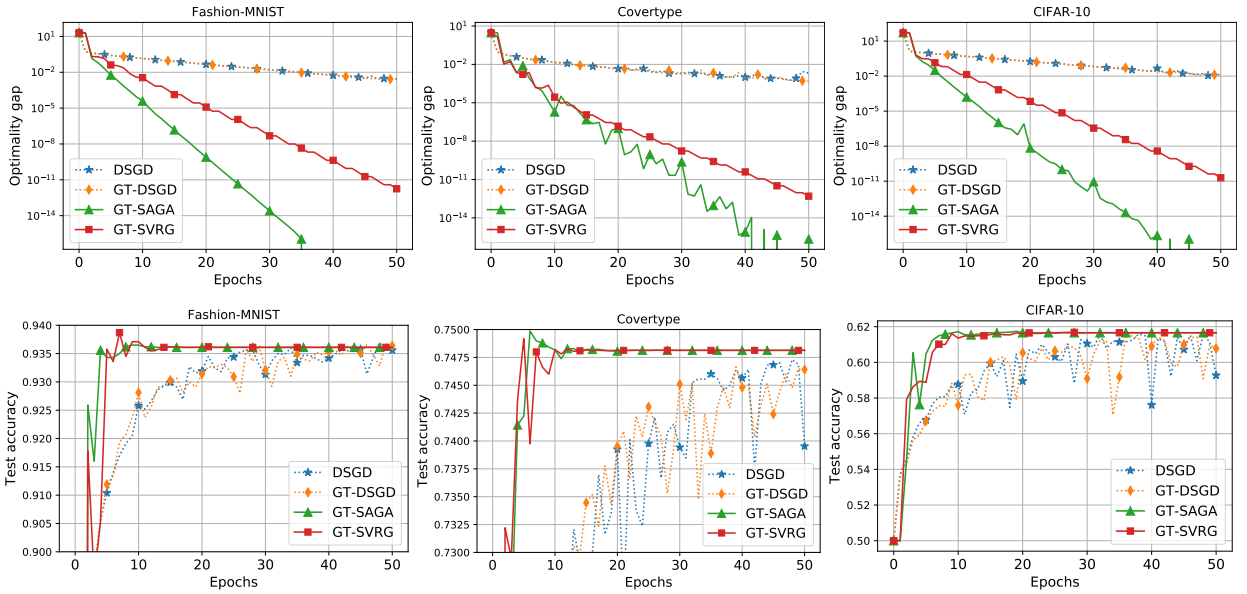

Figure 2.3: Performance comparison of **GT-SAGA** and **GT-SVRG** with **DSGD** and **GT-DSGD** on the directed exponential graph with $n = 10$ nodes over the Fashion-MNIST, Covertype, and CIFAR-10 datasets. The top row shows the optimality gap, while the bottom row shows the corresponding test accuracy.

GT-SAGA and **GT-SVRG** are not affected by the network topology. We next illustrate the speedup of **GT-SAGA** and **GT-SVRG** compared with their centralized counterparts. The speedup is characterized as the ratio of the number of component gradient computations required for centralized SAGA and SVRG that execute on *a single node* over the number of component gradient computations required *at each node* for **GT-SAGA** and **GT-SVRG** to achieve the optimality gap of 10^{-13}. It can be observed in Fig 2.2 (right) that linear speedup is achieved for both methods.

2.4.2 Comparison with the state-of-the-art

In this subsection, we compare the performances of the proposed **GT-SAGA** and **GT-SVRG** with the state-of-the-art decentralized stochastic first-order gradient algorithms over the datasets in Table 2.1, i.e., DSGD,

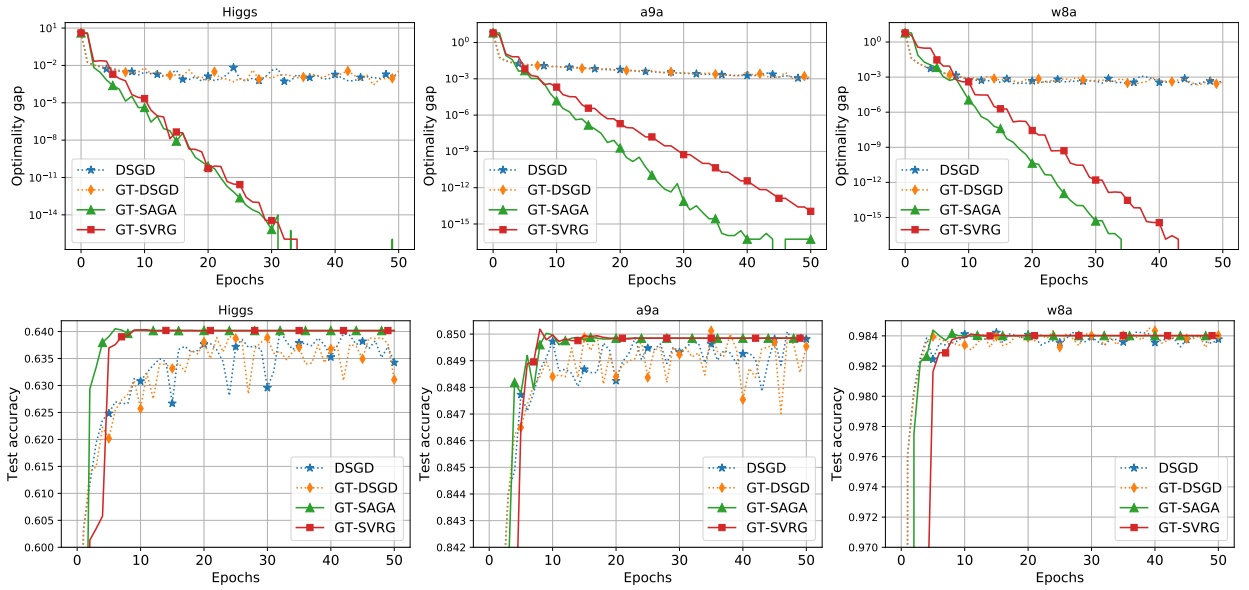

Figure 2.4: Performance comparison of **GT-SAGA** and **GT-SVRG** with **DSGD** and **GT-DSGD** on the directed exponential graph with $n = 10$ nodes over the Higgs, a9a, and w8a datasets. The top row presents the optimality gap, while the bottom row presents the corresponding test accuracy.

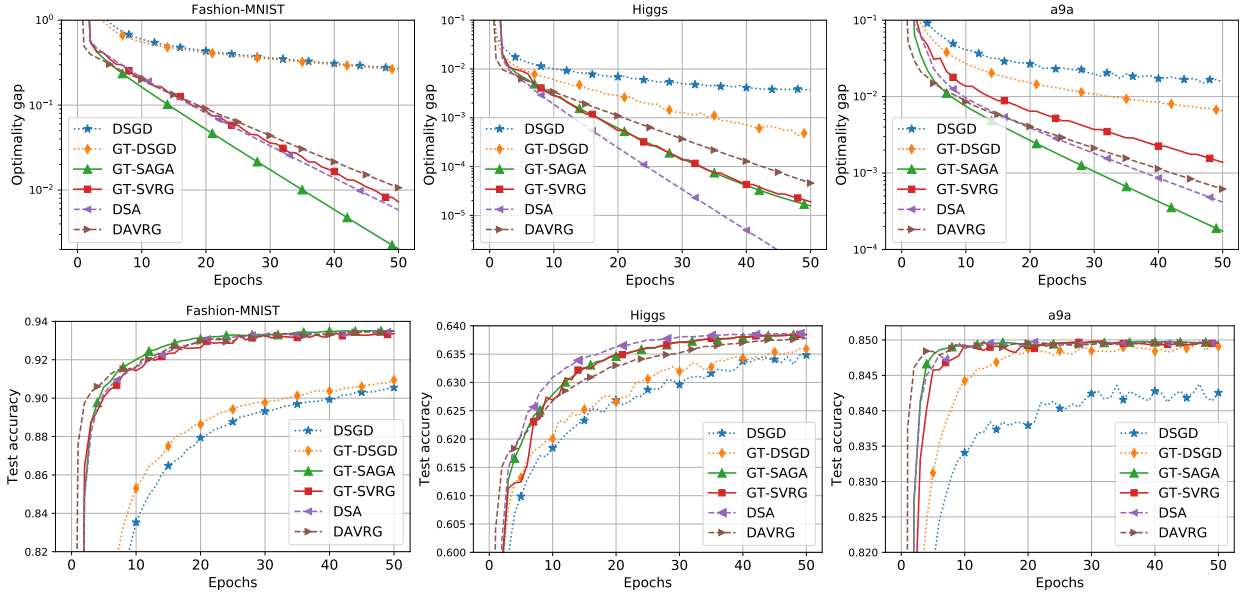

Figure 2.5: Comparison of **GT-SAGA** and **GT-SVRG** with **DSGD**, **GT-DSGD**, **DSA**, and **DAVRG** on an undirected nearest-neighbor geometric graph with $n = 200$ nodes over the Fashion-MNIST, Higgs, and a9a datasets. The top row shows the optimality gap, while the bottom row shows the corresponding test accuracy.

GT-DSGD, DSA, and DAVRG. We consider constant step-sizes for DSGD and GT-DSGD. Throughout this subsection, we set the regularization parameter as $\lambda = (nm)^{-1}$ for better test accuracy [23, 113].

We first consider the directed exponential graph with $n = 10$ nodes that typically arise e.g., in data centers [41] where data is divided among a small number of very well-connected nodes. Note that DSA and DAVRG are not applicable to directed graphs since they require symmetric weight matrices. We thus compare the performances of **GT-SAGA**, **GT-SVRG**, DSGD and GT-DSGD, presented in Figs. 2.3 and 2.4. It can be observed that the performances of DSGD and GT-DSGD are similar in this case, both of which linearly converge to a neighborhood of the optimal solution. On the other hand, **GT-SAGA** and **GT-SVRG** linearly converge to the *exact* optimal solution and, moreover, achieve better test accuracy faster.

We next consider a large-scale undirected geometric graph with $n = 200$ nodes that commonly arises e.g., in ad hoc network scenarios. The experimental result is presented in Fig. 2.5. We note that in this case GT-DSGD outperforms DSGD since the graph is not well-connected; this observation is consistent with [20, 31]. The performance of decentralized VR methods, **GT-SAGA**, **GT-SVRG**, DSA and DAVRG are rather comparable, all of which significantly outperform DSGD and GT-DSGD in terms of both optimality gap and test accuracy. However, we note that the theoretical guarantees of DSA and DAVRG are relatively weak, compared with that of **GT-SAGA** and **GT-SVRG**.

Finally, we observe that across all experiments shown in Figs. 2.3, 2.4, and 2.5, **GT-SAGA** exhibit faster convergence than **GT-SVRG**, at the expense of the storage cost of the gradient table at each node, demonstrating the space (storage) and time (convergence rate) tradeoffs of the SAGA and SVRG type variance reduction procedures.

2.5 Convergence analysis: A general dynamical system approach

Our goal is to develop a unified analysis framework for the **GT-VR** family of algorithms. To this aim, we first present a dynamical system that unifies the **GT-VR** algorithms and develop the results that can be used in general; see [52, 53, 56, 67] for similar approaches that do not involve local variance reduction schemes. Next, in Sections 2.5.3 and 2.5.4, we specialize this dynamical system for **GT-SAGA** and **GT-SVRG** in order to formally derive the main results of Section 2.3.

Recall that $\mathbf{x}_i^k \in \mathbb{R}^p$ denotes the **GT-VR** estimate of the optimal solution $\mathbf{x}^*$ at node i and iteration k, which iteratively descends in the direction of the global gradient tracker $\mathbf{y}_i^k \in \mathbb{R}^p$. Concatenating $\mathbf{x}_i^k$'s and $\mathbf{y}_i^k$'s in column vectors $\mathbf{x}^k, \mathbf{y}^k$, both in $\mathbb{R}^{pn}$, and defining $\mathbf{W} := \underline{W} \otimes \mathbf{I}_p$, we can write the estimate update of **GT-VR** as

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \alpha \mathbf{y}^k, \tag{2.4}$$

which is applicable to both **GT-SAGA** and **GT-SVRG**. The gradient tracking step next is given by

$$\mathbf{y}^{k+1} = \mathbf{W}\mathbf{y}^k + \mathbf{r}^{k+1} - \mathbf{r}^k, \tag{2.5}$$

where $\mathbf{r}^k \in \mathbb{R}^{pn}$ concatenates local variance-reduced gradient estimators $\mathbf{r}_i^k$'s, all in $\mathbb{R}^p$, which are given by $\mathbf{g}_i^k$'s in **GT-SAGA** and by $\mathbf{v}_i^k$'s in **GT-SVRG**. For the initial conditions, we have $\mathbf{y}^0 = \mathbf{r}^0 \in \mathbb{R}^p$ and $\mathbf{x}^0 \in \mathbb{R}^p$ is arbitrary.

Clearly, (2.4)-(2.5) are applicable to the **GT-VR** framework in general and the specialized algorithm of interest from this family can be obtained by using the corresponding variance-reduced estimator. We therefore first analyze the dynamical system (2.4)-(2.5), on top of which the specialized results for **GT-SAGA** and **GT-SVRG** are derived subsequently.

2.5.1 Preliminaries

To proceed, we define several auxiliary variables that will aid the subsequent convergence analysis as follows.

$$\begin{aligned}
\overline{\mathbf{x}}^k &\coloneqq \frac{1}{n}\left(\mathbf{1}_n^\top \otimes \mathbf{I}_p\right)\mathbf{x}^k,\\
\overline{\mathbf{y}}^k &\coloneqq \frac{1}{n}\left(\mathbf{1}_n^\top \otimes \mathbf{I}_p\right)\mathbf{y}^k,\\
\overline{\mathbf{r}}^k &\coloneqq \frac{1}{n}\left(\mathbf{1}_n^\top \otimes \mathbf{I}_p\right)\mathbf{r}^k,\\
\nabla\mathbf{f}(\mathbf{x}^k) &\coloneqq [\nabla f_1(\mathbf{x}_1^k)^\top, \ldots, \nabla f_n(\mathbf{x}_n^k)^\top]^\top,\\
\overline{\nabla\mathbf{f}}(\mathbf{x}^k) &\coloneqq \frac{1}{n}\left(\mathbf{1}_n^\top \otimes \mathbf{I}_p\right)\nabla\mathbf{f}(\mathbf{x}^k).
\end{aligned}$$

We recall that (2.5) is a stochastic gradient tracking method [4, 67, 126] as an application of dynamic consensus [51]. It is straightforward to verify by induction that [51]:

$$\overline{\mathbf{r}}^k = \overline{\mathbf{y}}^k, \qquad \forall k \geq 0.$$

Clearly, the randomness of both **GT-SAGA** and **GT-SVRG** lies in the set of independent random variables $\{s_i^k\}_{i=\{1,\cdots,n\}}^{k\geq 1}$. We denote $\mathcal{F}^k$ as the history of the dynamical system generated by $\{s_i^t\}_{i=\{1,\cdots,n\}}^{t\leq k-1}$. For both **GT-SAGA** and **GT-SVRG**, $\mathbf{r}_i^k$ is an unbiased estimator of $\nabla f_i(\mathbf{x}_i^k)$ given $\mathcal{F}^k$ [57, 61], i.e.,

$$\mathbb{E}\left[\mathbf{r}^k|\mathcal{F}^k\right] = \nabla\mathbf{f}(\mathbf{x}^k), \quad \mathbb{E}\left[\overline{\mathbf{y}}^k|\mathcal{F}^k\right] = \mathbb{E}\left[\overline{\mathbf{r}}^k|\mathcal{F}^k\right] = \overline{\nabla\mathbf{f}}(\mathbf{x}^k).$$

In the following, we first present a few well-known results related to decentralized gradient tracking methods whose proofs can be found in, e.g., [52–54, 56].

Lemma 2.5.1. *Let Assumptions 2.3.1 and 2.3.2 hold. If $0 < \alpha \leq \frac{1}{L}$, we have $\|\mathbf{x} - \alpha\nabla F(\mathbf{x}) - \mathbf{x}^*\| \leq (1-\mu\alpha)\|\mathbf{x}-\mathbf{x}^*\|$, $\forall \mathbf{x} \in \mathbb{R}^p$.*

Lemma 2.5.2. *Let Assumption 2.3.2 hold. Consider the iterates* $\{\mathbf{x}^k\}$ *generated by the dynamical system (2.4)-(2.5). We have that* $\left\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k) - \nabla F(\overline{\mathbf{x}}^k)\right\| \leq \frac{L}{\sqrt{n}} \left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|, \forall k \geq 0.$

Lemma 2.5.3. *Let Assumption 2.3.3 hold. We have that* $\forall \mathbf{x} \in \mathbb{R}^{np}$, $\|\mathbf{W}\mathbf{x} - \mathbf{J}\mathbf{x}\| \leq \lambda \|\mathbf{x} - \mathbf{J}\mathbf{x}\|$, *where* $\mathbf{J} = \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \otimes \mathbf{I}_p$.

2.5.2 Auxiliary results

In this subsection, we analyze the general dynamical system (2.4)-(2.5) by establishing the interrelationships between the mean-squared consensus error $\mathbb{E}\left[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right]$, network optimality gap $\mathbb{E}\left[\|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2\right]$ and gradient tracking error $\mathbb{E}\left[\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\|^2\right]$.

Lemma 2.5.4. *Let Assumption 2.3.3 hold. Consider the iterates* $\{\mathbf{x}^k\}$ *generated by (2.4)-(2.5). We have the following hold:* $\forall k \geq 0$,

$$\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\right\|^2\right] \leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + \frac{2\alpha^2}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right]. \tag{2.6}$$

$$\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\right\|^2\right] \leq 2\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + 2\alpha^2\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right]. \tag{2.7}$$

Proof. Using (2.4) and the fact that $\mathbf{J}\mathbf{W} = \mathbf{J}$, we have:

$$\left\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\right\|^2 = \left\|\mathbf{W}\mathbf{x}^k - \mathbf{J}\mathbf{x}^k - \alpha\left(\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right)\right\|^2 \tag{2.8}$$

Next, we use Young's inequality that $\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\eta)\|\mathbf{a}\|^2 + (1+\frac{1}{\eta})\|\mathbf{b}\|^2, \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^{np}, \forall \eta > 0$, and Lemma 2.5.3 in (2.8) to obtain: $\forall k \geq 0$,

$$\left\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\right\|^2 \leq (1+\eta)\lambda^2\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2 + \left(1+\eta^{-1}\right)\alpha^2\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2$$

Setting η as $\frac{1-\lambda^2}{2\lambda^2}$ and 1 in the above inequality respectively leads to (2.6) and (2.7). □

Next, we establish an inequality for $\mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right]$.

Lemma 2.5.5. *Let Assumptions 2.3.1, 2.3.2 and 2.3.3 hold. Consider the iterates* $\{\mathbf{x}^k\}$ *generated by (2.4)-(2.5). If* $0 < \alpha \leq \frac{1}{L}$, *we have the following inequalities hold:* $\forall k \geq 0$,

$$\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right] \leq \frac{L^2\alpha}{\mu}\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + (1-\mu\alpha)\mathbb{E}\left[n\|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2\right] + \frac{\alpha^2}{n}\mathbb{E}\left[\left\|\mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\|^2\right]. \tag{2.9}$$

$$\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right] \leq 2L^2\alpha^2\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + 2\mathbb{E}\left[n\|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2\right] + \frac{\alpha^2}{n}\mathbb{E}\left[\left\|\mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\|^2\right]. \tag{2.10}$$

Proof. Multiplying $\frac{\mathbf{1}_n^\top \otimes \mathbf{I}_p}{n}$ to (2.4), we have that $\forall k \geq 0$,

$$\overline{\mathbf{x}}^{k+1} = \overline{\mathbf{x}}^k - \alpha\overline{\mathbf{y}}^k = \overline{\mathbf{x}}^k - \alpha\overline{\mathbf{r}}^k.$$

We expand $\mathbb{E}\left[\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}^*\|^2|\mathcal{F}^k\right]$ as follows.

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}^*\right\|^2|\mathcal{F}^k\right] &= \mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k}-\alpha\overline{\mathbf{r}}^{k}-\mathbf{x}^*\right\|^2|\mathcal{F}^k\right]\\ &= \mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k}-\alpha\nabla F(\overline{\mathbf{x}}^k)-\mathbf{x}^*+\alpha\left(\nabla F(\overline{\mathbf{x}}^k)-\overline{\mathbf{r}}^k\right)\right\|^2|\mathcal{F}^k\right]\\ &= \left\|\overline{\mathbf{x}}^{k}-\alpha\nabla F(\overline{\mathbf{x}}^k)-\mathbf{x}^*\right\|^2+\alpha^2\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\mathbf{r}}^k|\right\|^2\mathcal{F}^k\right]\\ &\quad+2\alpha\Big\langle\overline{\mathbf{x}}^{k}-\alpha\nabla F(\overline{\mathbf{x}}^k)-\mathbf{x}^*,\nabla F(\overline{\mathbf{x}}^k)-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\Big\rangle,\end{aligned}\tag{2.11}$$

where in the last equality we used that $\mathbb{E}\left[\overline{\mathbf{r}}^k|\mathcal{F}^k\right]=\overline{\nabla\mathbf{f}}(\mathbf{x}^k)$. Next, we expand and simplify $\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\mathbf{r}}^k\right\|^2|\mathcal{F}^k\right]$:

$$\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\mathbf{r}}^k\right\|^2|\mathcal{F}^k\right]=\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\|^2+\mathbb{E}\left[\left\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)-\overline{\mathbf{r}}^k\right\|^2|\mathcal{F}^k\right]\tag{2.12}$$

where we used the fact that

$$\left\langle\nabla F(\overline{\mathbf{x}}^k)-\overline{\nabla\mathbf{f}}(\mathbf{x}^k),\mathbb{E}\left[\overline{\nabla\mathbf{f}}(\mathbf{x}^k)-\overline{\mathbf{r}}^k|\mathcal{F}^k\right]\right\rangle=0.$$

For the last term in (2.12), we have that:

$$\mathbb{E}\left[\left\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)-\overline{\mathbf{r}}^k\right\|^2|\mathcal{F}^k\right]=\frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n\left(\mathbf{r}_i^k-\nabla f_i(\mathbf{x}_i^k)\right)\right\|^2\Big|\mathcal{F}^k\right]=\frac{1}{n^2}\mathbb{E}\left[\left\|\mathbf{r}^k-\nabla\mathbf{f}(\mathbf{x}^k)\right\|^2|\mathcal{F}^k\right],\tag{2.13}$$

where in the equality above we used the fact that $\{\mathbf{r}_i^k\}_{i=1}^n$ are independent from each other and from $\mathcal{F}^k$ and therefore $\mathbb{E}\left[\sum_{i\neq j}\left\langle\mathbf{r}_i^k-\nabla f_i(\mathbf{x}_i^k),\mathbf{r}_j^k-\nabla f_j(\mathbf{x}_j^k)\right\rangle|\mathcal{F}^k\right]=0$. Now, we use (2.12), (2.13) and Lemma 2.5.1 in (2.11) to obtain:

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k+1}-\mathbf{x}^*\right\|^2|\mathcal{F}^k\right] &\leq (1-\mu\alpha)^2\|\overline{\mathbf{x}}^k-\mathbf{x}^*\|^2+\alpha^2\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\|^2\\ &\quad+2\alpha(1-\mu\alpha)\left\|\overline{\mathbf{x}}^k-\mathbf{x}^*\right\|\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\|\\ &\quad+\frac{\alpha^2}{n^2}\mathbb{E}\left[\left\|\mathbf{r}^k-\nabla\mathbf{f}(\mathbf{x}^k)\right\|^2|\mathcal{F}^k\right].\end{aligned}\tag{2.14}$$

Finally, we apply Young's inequality such that

$$2\alpha\left\|\overline{\mathbf{x}}^k-\mathbf{x}^*\right\|\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\|\leq\mu\alpha\left\|\overline{\mathbf{x}}^k-\mathbf{x}^*\right\|^2+\mu^{-1}\alpha\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\|^2$$

and $\left\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)-\nabla F(\overline{\mathbf{x}}^k)\right\|\leq\frac{L}{\sqrt{n}}\left\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\right\|,\forall k\geq0$, from Lemma 2.5.2 to (2.14) and take the total expectation; the resulting inequality is exactly (2.9). Similarly, using

$$2\alpha\left\|\overline{\mathbf{x}}^k-\mathbf{x}^*\right\|\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\|\leq\left\|\overline{\mathbf{x}}^k-\mathbf{x}^*\right\|^2+\alpha^2\left\|\nabla F(\overline{\mathbf{x}}^k)-\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\|$$

and Lemma 2.5.2 in (2.14) leads to (2.10). □

Next, we derive an inequality for $\mathbb{E}\left[\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2\right]$.

Lemma 2.5.6. *Let Assumption 2.3.2 and Assumption 2.3.3 hold. Consider the iterates $\{\mathbf{y}^k\}$ generated by (2.4)-(2.5). If $0 < \alpha \leq \frac{1}{4\sqrt{2}L}$, we have the following inequality hold: $\forall k \geq 0$,*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\right\|^2\right] \leq \frac{33L^2}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + \frac{L^2}{1-\lambda^2}\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] \\ + \left(\frac{1+\lambda^2}{2} + \frac{32L^2\alpha^2}{1-\lambda^2}\right)\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right] \\ + \frac{5}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\|^2\right] + \frac{4}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{r}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\right\|^2\right].\end{aligned}$$

Proof. Using (2.5) and the fact that $\mathbf{JW} = \mathbf{J}$, we have:

$$\begin{aligned}\left\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\right\|^2 &= \left\|\mathbf{W}\mathbf{y}^k + \mathbf{r}^{k+1} - \mathbf{r}^k - \mathbf{J}\left(\mathbf{W}\mathbf{y}^k + \mathbf{r}^{k+1} - \mathbf{r}^k\right)\right\|^2 \\ &= \left\|\mathbf{W}\mathbf{y}^k - \mathbf{J}\mathbf{y}^k + (\mathbf{I}_{np} - \mathbf{J})\left(\mathbf{r}^{k+1} - \mathbf{r}^k\right)\right\|^2. \end{aligned} \tag{2.15}$$

To proceed from (2.15), we use Young's inequality that $\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\eta)\|\mathbf{a}\|^2 + (1+\frac{1}{\eta})\|\mathbf{b}\|^2, \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^{np}$ with $\eta = \frac{2\lambda^2}{1-\lambda^2}$ and that $\|\!|\mathbf{I}_{np} - \mathbf{J}|\!\| = 1$ together with Lemma 2.5.3 to obtain:

$$\begin{aligned}\left\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\right\|^2 &\leq \left(1 + \frac{1-\lambda^2}{2\lambda^2}\right)\left\|\mathbf{W}\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2 + \left(1 + \frac{2\lambda^2}{1-\lambda^2}\right)\left\|(\mathbf{I}_{np} - \mathbf{J})\left(\mathbf{r}^{k+1} - \mathbf{r}^k\right)\right\|^2 \\ &\leq \frac{1+\lambda^2}{2}\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2 + \frac{2}{1-\lambda^2}\left\|\mathbf{r}^{k+1} - \mathbf{r}^k\right\|^2. \end{aligned} \tag{2.16}$$

We then take the total expectation to obtain:

$$\mathbb{E}\left[\left\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\right\|^2\right] \leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right] + \frac{2}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{r}^{k+1} - \mathbf{r}^k\right\|^2\right] \tag{2.17}$$

Now, we derive an upper bound for $\mathbb{E}[\|\mathbf{r}^{k+1} - \mathbf{r}^k\|^2]$. Firstly,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{r}^{k+1} - \mathbf{r}^k\|^2\right] &\leq 2\mathbb{E}\left[\|\mathbf{r}^{k+1} - \mathbf{r}^k - (\nabla\mathbf{f}(\mathbf{x}^{k+1}) - \nabla\mathbf{f}(\mathbf{x}^k))\|^2\right] + 2\mathbb{E}\left[\|\nabla\mathbf{f}(\mathbf{x}^{k+1}) - \nabla\mathbf{f}(\mathbf{x}^k)\|^2\right] \\ &\leq 2\mathbb{E}\left[\|\mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\|^2\right] + 2\mathbb{E}\left[\|\mathbf{r}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2\right] + 2L^2\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2\right] \end{aligned} \tag{2.18}$$

where in the last inequality above we used that

$$\mathbb{E}\left[\langle \mathbf{r}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1}), \mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{r}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1}), \mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\rangle | \mathcal{F}^{k+1}\right]\right] = 0.$$

We next bound $\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right]$. Using (2.4) leads to:

$$\begin{aligned}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2 &= \left\|\mathbf{W}\mathbf{x}^k - \alpha\mathbf{y}^k - \mathbf{x}^k\right\|^2 \\ &= \left\|(\mathbf{W} - \mathbf{I}_{np})\left(\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right) - \alpha\mathbf{y}^k\right\|^2 \\ &\leq 8\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2 + 2\alpha^2\left\|\mathbf{y}^k\right\|^2, \end{aligned} \tag{2.19}$$

where in (2.19) we used the fact that $\|\!|\mathbf{W}-\mathbf{I}_{np}|\!\| \leq 2$. We then denote $\nabla\mathbf{f}(\mathbf{x}^*) \coloneqq \left[\nabla f_1(\mathbf{x}^*)^\top, \cdots, \nabla f_n(\mathbf{x}^*)^\top\right]^\top$ and note that $(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\nabla\mathbf{f}(\mathbf{x}^*) = \mathbf{0}_p$. We bound $\|\mathbf{y}^k\|$ as follows.

$$\begin{aligned}\left\|\mathbf{y}^k\right\| =& \left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k + \mathbf{J}\mathbf{r}^k - \mathbf{J}\nabla\mathbf{f}(\mathbf{x}^k) + \mathbf{J}\nabla\mathbf{f}(\mathbf{x}^k) - \mathbf{J}\nabla\mathbf{f}(\mathbf{x}^*)\right\| \\ \leq& \left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\| + \left\|\mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\| + L\left\|\mathbf{x}^k - (\mathbf{1}_n \otimes \mathbf{I}_p)\mathbf{x}^*\right\| \\ \leq& \left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\| + \left\|\mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\| + L\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\| + \sqrt{n}L\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|,\end{aligned}$$

where in the first equality we used $\overline{\mathbf{y}}^k = \overline{\mathbf{r}}^k, \forall k \geq 0$. Squaring the above inequality obtains the following:

$$\left\|\mathbf{y}^k\right\|^2 \leq 4L^2\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2 + 4nL^2\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2 + 4\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2 + 4\left\|\mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\|^2. \tag{2.20}$$

Using (2.20) in (2.19) with the requirement that $0 < \alpha \leq \frac{1}{4\sqrt{2}L}$ and taking the total expectation, we have:

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] \leq& 8.25\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + 0.25\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] \\ &+ 8\alpha^2\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right] + 8\alpha^2\mathbb{E}\left[\left\|\mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\|^2\right].\end{aligned} \tag{2.21}$$

Finally, we apply (2.21) in (2.18) with $0 < \alpha \leq \frac{1}{4\sqrt{2}L}$ to obtain:

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{r}^{k+1} - \mathbf{r}^k\right\|^2\right] \leq& 16.5L^2\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + 0.5L^2\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] \\ &+ 16\alpha^2L^2\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right] + 2.5\mathbb{E}\left[\left\|\mathbf{r}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\|^2\right] + 2\mathbb{E}\left[\left\|\mathbf{r}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\right\|^2\right].\end{aligned}$$

Using the above inequality in (2.17) completes the proof. □

We finally present a general convergence result on a sequence of random variables that converge linearly in the mean-square sense. We note that this result is implied in the probability literature; see [127] for example. For the sake of completeness, we present its proof here.

Lemma 2.5.7. *Let $\{X_k\}_{k\geq 0}$ be a sequence of random variables such that $\mathbb{E}[|X_k|] \leq \gamma^k$ for some $0 < \gamma < 1$. Then we have*

$$\mathbb{P}\left(\lim_{k\to\infty}(\gamma+\delta)^{-k}|X_k| = 0\right) = 1,$$

where $\delta > 0$ is an arbitrary positive constant.

Proof. By Chebyshev's inequality, we have: $\forall\epsilon > 0, \forall\delta > 0$,

$$\begin{aligned}\mathbb{P}\left((\gamma+\delta)^{-k}|X_k| > \epsilon\right) \leq& \epsilon^{-1}\mathbb{E}[(\gamma+\delta)^{-k}|X_k|] \\ \leq& \epsilon^{-1}(\gamma/(\gamma+\delta))^k.\end{aligned}$$

Summing the inequality above over k, we obtain:

$$\sum_{k=0}^{\infty} \mathbb{P}\left((\gamma+\delta)^{-k}|X_k| > \epsilon\right) \leq \epsilon^{-1} \sum_{k=0}^{\infty} \left(\frac{\gamma}{\gamma+\delta}\right)^k < \infty.$$

By the Borel-Cantelli lemma,

$$\mathbb{P}\left((\gamma+\delta)^{-k}|X_k| > \epsilon \text{ for infinitely many } k\right) = 0,$$

and the proof follows. □

We note that Lemma 2.5.7 states that the non-asymptotic linear convergence of a sequence of random variables in the mean-square sense implies its asymptotic linear convergence in the almost sure sense. As a consequence, Corollaries 2.3.1 and 2.3.2 will be immediately at hand once Theorems 2.3.1 and 2.3.2 are established. With the help of the auxiliary results on the general dynamical system (2.4)-(2.5) established in this section, we now derive explicit convergence rates for the proposed algorithms, **GT-SAGA** and **GT-SVRG**, in the next sections.

2.5.3 Analysis of GT-SAGA

In this section, we establish the mean-square linear convergence of **GT-SAGA** described in Algorithm 1. Following the unified representation in (2.4)-(2.5), we recall that the local gradient estimator $\mathbf{r}_i^k$ is given by $\mathbf{g}_i^k$ in **GT-SAGA**, i.e., $\forall i \in \{1, \cdots, n\}, \forall k \geq 1$,

$$\mathbf{g}_i^k = \nabla f_{i,s_i^k}(\mathbf{x}_i^k) - \nabla f_{i,s_i^k}(\mathbf{z}_{i,s_i^k}^k) + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla f_{i,j}(\mathbf{z}_{i,j}^k),$$

where s_i^k is selected uniformly at random from $\{1, \cdots, m_i\}$ and the auxiliary variable $\mathbf{z}_{i,j}^k$ is the most recent iterate where the component gradient $\nabla f_{i,j}$ was evaluated up to time k.

2.5.3.1 Bounding the variance of the gradient estimator

We first derive an upper bound for $\mathbb{E}\left[\|\mathbf{g}^k - \nabla \mathbf{f}(\mathbf{x}^k)\|^2\right]$ that is the variance of the gradient estimator $\mathbf{g}^k$. To do this, we define Υ_i^k as the averaged optimality gap of the auxiliary variables of $\{\mathbf{z}_{i,j}^k\}_{j=1}^{m_i}$ at node i:

$$\Upsilon_i^k := \frac{1}{m_i}\sum_{j=1}^{m_i} \left\|\mathbf{z}_{i,j}^k - \mathbf{x}^*\right\|^2, \qquad \Upsilon^k := \sum_{i=1}^{n} \Upsilon_i^k. \tag{2.22}$$

The next lemma shows that Υ^k admits an intrinsic contraction. Recall that $M = \max_i m_i$ and $m = \min_i m_i$.

Lemma 2.5.8. *Consider the iterates $\{\Upsilon^k\}$ generated by **GT-SAGA**. We have the following holds: $\forall k \geq 1$,*

$$\mathbb{E}\left[\Upsilon^{k+1}\right] \leq \left(1 - \frac{1}{M}\right)\mathbb{E}\left[\Upsilon^k\right] + \frac{2}{m}\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + \frac{2}{m}\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right].$$

Proof. Recall Algorithm 1 and note that $\forall k \geq 1$, $\mathbf{z}_{i,j}^{k+1} = \mathbf{z}_{i,j}^{k}$ with probability $1 - \frac{1}{m_i}$ and $\mathbf{z}_{i,j}^{k+1} = \mathbf{x}_i^k$ with probability $\frac{1}{m_i}$ given $\mathcal{F}^k$. Then we have the following holds: $\forall i, \forall k \geq 1$,

$$\begin{aligned}
&\mathbb{E}\left[\Upsilon_i^{k+1}|\mathcal{F}^k\right] \\
&= \frac{1}{m_i}\sum_{j=1}^{m_i}\mathbb{E}\left[\left\|\mathbf{z}_{i,j}^{k+1} - \mathbf{x}^*\right\|^2|\mathcal{F}^k\right] \\
&= \frac{1}{m_i}\sum_{j=1}^{m_i}\mathbb{E}\left[\left(1-\frac{1}{m_i}\right)\left\|\mathbf{z}_{i,j}^{k} - \mathbf{x}^*\right\|^2 + \frac{1}{m_i}\left\|\mathbf{x}_i^{k} - \mathbf{x}^*\right\|^2\Big|\mathcal{F}^k\right] \\
&= \left(1-\frac{1}{m_i}\right)\Upsilon_i^k + \frac{1}{m_i}\left\|\mathbf{x}_i^{k} - \mathbf{x}^*\right\|^2 \\
&\leq \left(1-\frac{1}{M}\right)\Upsilon_i^k + \frac{2}{m}\left\|\mathbf{x}_i^{k} - \overline{\mathbf{x}}^k\right\|^2 + \frac{2}{m}\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2
\end{aligned} \tag{2.23}$$

The proof follows by summing (2.23) over i and taking the total expectation. □

In the next lemma, we bound the stochastic gradient variance $\mathbb{E}\left[\|\mathbf{g}^k - \nabla\mathbf{f}(\mathbf{x}^k)\|^2\right]$ by the mean-square consensus error and the optimality gap of $\mathbf{x}^k$ and Υ^k.

Lemma 2.5.9. *Let Assumption 2.3.2 hold. Consider the iterates $\{\mathbf{g}^k\}$ generated by* ***GT-SAGA****. Then we have the following inequality hold: $\forall k \geq 1$,*

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\|^2\right] \leq 4L^2\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] + 2L^2\mathbb{E}\left[\Upsilon^k\right].$$

Proof. Recall the local gradient estimator $\mathbf{g}_i^k$ from Algorithm 1 and proceed as follows.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{g}_i^k - \nabla f_i(\mathbf{x}_i^k)\right\|^2|\mathcal{F}^k\right] \\
&= \mathbb{E}\Big[\Big\|\nabla f_{i,s_i^k}\left(\mathbf{x}_i^k\right) - \nabla f_{i,s_i^k}\left(\mathbf{z}_{i,s_i^k}^k\right) - \left(\nabla f_i(\mathbf{x}_i^k) - \tfrac{1}{m_i}\textstyle\sum_{j=1}^{m_i}\nabla f_{i,j}\left(\mathbf{z}_{i,j}^k\right)\right)\Big\|^2\Big|\mathcal{F}^k\Big] \\
&\leq \mathbb{E}\left[\left\|\nabla f_{i,s_i^k}\left(\mathbf{x}_i^k\right) - \nabla f_{i,s_i^k}\left(\mathbf{z}_{i,s_i^k}^k\right)\right\|^2\Big|\mathcal{F}^k\right] \\
&= \tfrac{1}{m_i}\textstyle\sum_{j=1}^{m_i}\left\|\left(\nabla f_{i,j}\left(\mathbf{x}_i^k\right) - \nabla f_{i,j}(\mathbf{x}^*)\right) + \left(\nabla f_{i,j}(\mathbf{x}^*) - \nabla f_{i,j}\left(\mathbf{z}_{i,j}^k\right)\right)\right\|^2 \\
&\leq 2L^2\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|^2 + 2L^2\Upsilon_i^k \\
&\leq 4L^2\left\|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\right\|^2 + 4L^2\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2 + 2L^2\Upsilon_i^k,
\end{aligned} \tag{2.24}$$

where the second inequality uses the standard conditional variance decomposition

$$\mathbb{E}\left[\left\|\mathbf{a}_i^k - \mathbb{E}\left[\mathbf{a}_i^k|\mathcal{F}^k\right]\right\|^2|\mathcal{F}^k\right] = \mathbb{E}\left[\left\|\mathbf{a}_i^k\right\|^2|\mathcal{F}^k\right] - \left\|\mathbb{E}\left[\mathbf{a}_i^k|\mathcal{F}^k\right]\right\|^2 \leq \mathbb{E}\left[\left\|\mathbf{a}_i^k\right\|^2|\mathcal{F}^k\right], \tag{2.25}$$

with $\mathbf{a}_i^k = \nabla f_{i,s_i^k}\left(\mathbf{x}_i^k\right) - \nabla f_{i,s_i^k}\left(\mathbf{z}_{i,s_i^k}^k\right)$. The proof follows by summing (2.24) over i. □

Lemma 2.5.9 clearly shows that as $\mathbf{x}_i^k$ and $\mathbf{z}_{i,j}^k$ approach to an agreement on $\mathbf{x}^*$, the variance of the gradient estimator decays to zero. We have the following corollary.

Corollary 2.5.1. *Let Assumption 2.3.2 and 2.3.3 hold. Consider the iterates* $\{\mathbf{g}^k\}$ *generated by* **GT-SAGA**. *If* $0 < \alpha \leq \frac{1}{4\sqrt{2}L}$, *then the following inequality holds* $\forall k \geq 0$,

$$\mathbb{E}\left[\left\|\mathbf{g}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\right\|^2\right] \leq 12.75L^2\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + 12.5L^2\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] + 8L^2\alpha^2\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right] + 2.25L^2\mathbb{E}\left[t^k\right].$$

Proof. Following directly from Lemma 2.5.9, we have: $\forall k \geq 0$,

$$\mathbb{E}\left[\left\|\mathbf{g}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\right\|^2\right] \leq 4L^2\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right] + 2L^2\mathbb{E}\left[t^{k+1}\right].$$

Using (2.7), (2.10) and Lemma 2.5.8 in the inequality above leads to the following: if $0 < \alpha \leq \frac{1}{4\sqrt{2}L}$,

$$\mathbb{E}\left[\left\|\mathbf{g}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\right\|^2\right] \leq 12.25L^2\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + 12L^2\mathbb{E}\left[n\|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2\right] + 8L^2\alpha^2\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right] + 2L^2\mathbb{E}\left[t^k\right] + 0.125\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\|^2\right].$$

The proof follows by applying Lemma 2.5.9 in the above. □

2.5.3.2 Proof of Theorem 2.3.1

With the bounds on the gradient variance for **GT-SAGA** derived in the previous subsection, we are now able to refine the inequalities obtained for the general dynamical system (2.4)-(2.5) in Section 2.5 and derive the explicit convergence rates for **GT-SAGA**. First, we apply the upper bound on $\mathbb{E}[\|\mathbf{g}^k - \nabla\mathbf{f}(\mathbf{x}^k)\|^2]$ in Lemma 2.5.9 to (2.9) to obtain: $\forall k \geq 0$,

$$\begin{aligned}&\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right]\\ &\leq L^2\alpha\left(\frac{1}{\mu} + \frac{4\alpha}{n}\right)\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + \left(1 - \mu\alpha + \frac{4L^2\alpha^2}{n}\right)\mathbb{E}\left[n\|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2\right] + \frac{2L^2\alpha^2}{n}\mathbb{E}\left[t^k\right].\end{aligned}$$

If $0 < \alpha \leq \frac{1}{4\mu}$, then $\frac{1}{\mu} + \frac{4\alpha}{n} \leq \frac{2}{\mu}$; if $0 < \alpha \leq \frac{\mu n}{8L^2}$, then we have $1 - \mu\alpha + \frac{4L^2\alpha^2}{n} \leq 1 - \frac{\mu\alpha}{2}$. Therefore, if $0 < \alpha \leq \frac{\mu}{8L^2}$, we have the following: $\forall k \geq 0$,

$$\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right] \leq \frac{2L^2\alpha}{\mu}\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + \left(1 - \frac{\mu\alpha}{2}\right)\mathbb{E}\left[n\|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2\right] + \frac{2L^2\alpha^2}{n}\mathbb{E}\left[t^k\right] \tag{2.26}$$

Second, we apply the upper bounds on $\mathbb{E}[\|\mathbf{g}^k - \nabla\mathbf{f}(\mathbf{x}^k)\|^2]$ and $\mathbb{E}[\|\mathbf{g}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2]$ in Lemma 2.5.9 and Corollary 2.5.1 to Lemma 2.5.6 to obtain the following: $\forall k \geq 0$,

$$\mathbb{E}\left[\left\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\right\|^2\right] \leq \frac{104L^2}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + \frac{71L^2}{1-\lambda^2}\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] + \frac{19L^2}{1-\lambda^2}\mathbb{E}\left[t^k\right] + \frac{3+\lambda^2}{4}\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right], \tag{2.27}$$

if $0 < \alpha \leq \frac{1-\lambda^2}{16L}$. We next write (2.6), (2.26), Lemma 2.5.8 and (2.27) jointly as a linear matrix inequality.

Proposition 2.5.1. *Let Assumptions 2.3.1, 2.3.2, 2.3.3 hold and consider the iterates* $\{\mathbf{x}^k\},\{\mathbf{y}^k\},\{\Upsilon^k\}$ *generated by* **GT-SAGA**. *If the step-size* α *follows* $0<\alpha\leq\frac{\mu(1-\lambda)}{16L^2}$, *we have:* $\forall k\geq 1$,

$$\mathbf{u}^{k+1}\leq \mathbf{G}_\alpha \mathbf{u}^k, \tag{2.28}$$

where $\mathbf{u}^k\in\mathbb{R}^4$ *and* $\mathbf{G}_\alpha\in\mathbb{R}^{4\times 4}$ *are defined as follows:*

$$\mathbf{u}^k=\begin{bmatrix}\mathbb{E}\left[\left\|\mathbf{x}^k-\mathbf{J}\mathbf{x}^k\right\|^2\right]\\ \mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k-\mathbf{x}^*\right\|^2\right]\\ \mathbb{E}\left[\Upsilon^k\right]\\ \mathbb{E}\left[L^{-2}\left\|\mathbf{y}^k-\mathbf{J}\mathbf{y}^k\right\|^2\right]\end{bmatrix},\qquad \mathbf{G}_\alpha=\begin{bmatrix}\frac{1+\lambda^2}{2} & 0 & 0 & \frac{2\alpha^2L^2}{1-\lambda^2}\\ \frac{2L^2\alpha}{\mu} & 1-\frac{\mu\alpha}{2} & \frac{2L^2\alpha^2}{n} & 0\\ \frac{2}{m} & \frac{2}{m} & 1-\frac{1}{M} & 0\\ \frac{104}{1-\lambda^2} & \frac{71}{1-\lambda^2} & \frac{19}{1-\lambda^2} & \frac{3+\lambda^2}{4}\end{bmatrix}.$$

Clearly, to show the linear convergence of **GT-SAGA**, it suffices to derive the range of α such that $\rho(\mathbf{G}_\alpha)<1$. To do this, we present a useful lemma from [36].

Lemma 2.5.10. *Let* $\mathbf{A}\in\mathbb{R}^{d\times d}$ *be non-negative and* $\mathbf{x}\in\mathbb{R}^d$ *be positive. If* $\mathbf{A}\mathbf{x}\leq\beta\mathbf{x}$ *for* $\beta>0$, *then* $\rho(\mathbf{A})\leq\left\|\left|\mathbf{A}\right|\right\|_\infty^{\mathbf{x}}\leq\beta$.

We are ready to prove Theorem 2.3.1 based on Proposition 2.5.1.

Proof of Theorem 2.3.1. Recall from Proposition 2.5.1 that if $0<\alpha\leq\frac{\mu(1-\lambda)}{16L^2}$, we have $\mathbf{u}^{k+1}\leq\mathbf{G}_\alpha\mathbf{u}^k$. In the light of Lemma 2.5.10, we solve for the range of the step-size α and a positive vector $\boldsymbol{\epsilon}=[\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4]^\top$ such that the following (entry-wise) linear matrix inequality holds:

$$\mathbf{G}_\alpha\boldsymbol{\epsilon}\leq\left(1-\frac{\mu\alpha}{4}\right)\boldsymbol{\epsilon}, \tag{2.29}$$

which can be written equivalently in the following form:

$$\frac{\mu\alpha}{4}+\frac{2L^2}{1-\lambda^2}\frac{\epsilon_4}{\epsilon_1}\alpha^2\leq\frac{1-\lambda^2}{2} \tag{2.30}$$

$$\frac{2L^2}{n}\epsilon_3\alpha\leq\frac{\mu}{4}\epsilon_2-\frac{2L^2}{\mu}\epsilon_1 \tag{2.31}$$

$$\frac{\mu\alpha}{4}\leq\frac{1}{M}-\frac{2}{m}\frac{\epsilon_1}{\epsilon_3}-\frac{2}{m}\frac{\epsilon_2}{\epsilon_3} \tag{2.32}$$

$$\frac{\mu\alpha}{4}\leq\frac{1-\lambda^2}{4}-\frac{104}{1-\lambda^2}\frac{\epsilon_1}{\epsilon_4}-\frac{71}{1-\lambda^2}\frac{\epsilon_2}{\epsilon_4}-\frac{19}{1-\lambda^2}\frac{\epsilon_3}{\epsilon_4} \tag{2.33}$$

Clearly, that (2.31)–(2.33) hold for some feasible range of α is equivalent to the RHS of (2.31)–(2.33) being positive. Based on this observation, we will next fix the values of $\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4$ that are independent of α. First, for the RHS of (2.31) to be positive, we set $\epsilon_1=1,\epsilon_2=8.5Q^2$, where $Q=L/\mu$. Second, the RHS

of (2.32) being positive is equivalent to

$$\epsilon_3 > \frac{2M}{m}\epsilon_1 + \frac{2M}{m}\epsilon_2 = \frac{2M}{m} + \frac{17MQ^2}{m}. \tag{2.34}$$

We therefore set $\epsilon_3 = \frac{20MQ^2}{m}$. Third, we note that the RHS of (2.33) being positive is equivalent to the following:

$$\begin{aligned}\epsilon_4 &> \frac{4}{(1-\lambda^2)^2}\left(104\epsilon_1 + 71\epsilon_2 + 19\epsilon_3\right)\\ &= \frac{4}{(1-\lambda^2)^2}\left(104 + 603.5Q^2 + \frac{380MQ^2}{m}\right)\end{aligned}$$

Note that $104 + 603.5Q^2 + \frac{380MQ^2}{m} \leq \frac{1087.5MQ^2}{m}$. We therefore set $\epsilon_4 = \frac{8700}{(1-\lambda^2)^2}\frac{MQ^2}{m}$. We now solve for the range of α from (2.30)–(2.33) given the previously fixed $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$. From (2.31), we have that

$$\alpha \leq \frac{n}{2L^2\epsilon_3}\left(\frac{\mu}{4}\epsilon_2 - \frac{2L^2}{\mu}\epsilon_1\right) = \frac{m}{M}\frac{n}{320QL}. \tag{2.35}$$

Moreover, it is straightforward to verify that if α satisfies

$$0 < \alpha \leq \frac{m}{M}\frac{(1-\lambda^2)^2}{320QL} \tag{2.36}$$

then (2.30) holds. Next, to make (2.32) hold, it suffices to make α:

$$\alpha \leq \frac{1}{5\mu M}. \tag{2.37}$$

Finally, to make (2.33) hold, it suffices to make

$$\alpha \leq \frac{1-\lambda^2}{2\mu}. \tag{2.38}$$

To summarize, combining (2.36)–(2.38), we conclude that if the step-size α satisfies

$$0 < \alpha \leq \overline{\alpha} := \min\left\{\frac{1}{5\mu M}, \frac{m}{320M}\frac{(1-\lambda^2)^2}{LQ}\right\}, \tag{2.39}$$

then (2.29) holds with some $\boldsymbol{\epsilon} > 0$ and thus $\rho\left(\mathbf{G}_\alpha\right) \leq 1 - \frac{\mu\alpha}{4}$ according to Lemma 2.5.10. Furhter if $\alpha = \overline{\alpha}$, we have

$$\rho\left(\mathbf{G}_\alpha\right) \leq 1 - \min\left\{\frac{1}{20M}, \frac{m}{1280M}\frac{(1-\lambda^2)^2}{Q^2}\right\},$$

which completes the proof. □

2.5.4 Analysis of `GT-SVRG`

In this section, we conduct the complexity analysis of **`GT-SVRG`** in Algorithm 2 based on the auxiliary results derived for the general dynamical system (2.4)-(2.5) in Section 2.5. Recall from Algorithm 2 that the gradient

estimator $\mathbf{v}_i^k$ at each node i in **GT-SVRG** is given by the following: $\forall k \geq 1$, choose s_i^k uniformly at random in $\{1, \cdots, m_i\}$ and

$$\mathbf{v}_i^k = \nabla f_{i,s_i^k}(\mathbf{x}_i^k) - \nabla f_{i,s_i^k}(\boldsymbol{\tau}_i^k) + \nabla f_i(\boldsymbol{\tau}_i^k) \tag{2.40}$$

where $\boldsymbol{\tau}_i^k = \mathbf{x}_i^k$ if $\mathrm{mod}(k, T) = 0$, where T is the length of each inner loop iterations of **GT-SVRG**; otherwise $\boldsymbol{\tau}_i^k = \boldsymbol{\tau}_i^{k-1}$. To proceed, we define an auxiliary variable $\overline{\boldsymbol{\tau}}^k := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\tau}_i^k$, $\forall k \geq 0$.

2.5.4.1 Bounding the variance of the gradient estimator

We first bound the variance of $\mathbf{v}_i^k$, following a similar procedure as the proof of Lemma 2.5.9.

Lemma 2.5.11. *Let Assumption 2.3.2 hold and consider the iterates $\{\mathbf{v}^k\}$ generated by **GT-SVRG** in Algorithm 2. The following inequality holds $\forall k \geq 0$:*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{v}^k - \nabla\mathbf{f}(\mathbf{x}^k)\right\|^2\right] \leq\ & 4L^2\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] \\ & + 4L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^k - \mathbf{J}\boldsymbol{\tau}^k\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^k - \mathbf{x}^*\right\|^2\right].\end{aligned}$$

Proof. We recall from Algorithm 2 the definition of $\mathbf{v}_i^k$ in **GT-SVRG** and proceed as follows.

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{v}_i^k - \nabla f_i(\mathbf{x}_i^k)\right\|^2 |\mathcal{F}^k\right] \\ &= \mathbb{E}\Big[\Big\|\nabla f_{i,s_i^k}(\mathbf{x}_i^k) - \nabla f_{i,s_i^k}(\boldsymbol{\tau}_i^k) - \left(\nabla f_i(\mathbf{x}_i^k) - \nabla f_i(\boldsymbol{\tau}_i^k)\right)\Big\|^2\Big|\mathcal{F}^k\Big] \\ &\leq \mathbb{E}\left[\left\|\nabla f_{i,s_i^k}(\mathbf{x}_i^k) - \nabla f_{i,s_i^k}(\boldsymbol{\tau}_i^k)\right\|^2\Big|\mathcal{F}^k\right] \\ &= \frac{1}{m_i}\sum_{j=1}^{m_i}\Big\|\left(\nabla f_{i,j}(\mathbf{x}_i^k) - \nabla f_{i,j}(\mathbf{x}^*)\right) + \left(\nabla f_{i,j}(\mathbf{x}^*) - \nabla f_{i,j}(\boldsymbol{\tau}_i^k)\right)\Big\|^2 \\ &\leq 2L^2\left\|\mathbf{x}_i^k - \mathbf{x}^*\right\|^2 + 2L^2\left\|\boldsymbol{\tau}_i^k - \mathbf{x}^*\right\|^2 \\ &\leq 4L^2\left\|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\right\|^2 + 4L^2\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2 + 4L^2\left\|\boldsymbol{\tau}_i^k - \overline{\boldsymbol{\tau}}^k\right\|^2 + 4L^2\left\|\overline{\boldsymbol{\tau}}^k - \mathbf{x}^*\right\|^2,\end{aligned} \tag{2.41}$$

where in the second inequality we used the standard conditional variance decomposition in (2.25). The proof follows by summing (2.41) over i and taking the total expectation. □

Lemma 2.5.11 shows that as $\mathbf{x}^k$ and $\boldsymbol{\tau}^k$ progressively approach the optimal solution $\mathbf{x}^*$ of the Problem (2.1), the variance of the gradient estimator $\mathbf{v}^k$ goes to zero. We then have the following corollary.

Corollary 2.5.2. *Let Assumption 2.3.2 hold and consider the iterates $\{\mathbf{v}^k\}$ generated by **GT-SVRG**. If $0 < \alpha \leq \frac{1}{8L}$, then the following inequality holds $\forall k \geq 0$:*

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{v}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\right\|^2\right] \leq\ & 16.75L^2\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + 16L^2\alpha^2\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right] + 16.5L^2\mathbb{E}\left[n\|\overline{\mathbf{x}}^k - \mathbf{x}^*\|^2\right] \\ & + 4.5L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^k - \mathbf{J}\boldsymbol{\tau}^k\right\|^2\right] + 4.5L^2\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^k - \mathbf{x}^*\right\|^2\right].\end{aligned}$$

Proof. From Lemma 2.5.11, we have: $\forall k \geq 0$,

$$\mathbb{E}\left[\left\|\mathbf{v}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\right\|^2\right] \leq 4L^2\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right] + 4L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^{k+1} - \mathbf{J}\boldsymbol{\tau}^{k+1}\right\|^2\right] + 4L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^{k+1} - \mathbf{J}\boldsymbol{\tau}^{k+1}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^{k+1} - \mathbf{x}^*\right\|^2\right]. \tag{2.42}$$

Recall that $\boldsymbol{\tau}^{k+1} = \mathbf{x}^{k+1}$ if $\mathrm{mod}(k+1, T) = 0$; otherwise, $\boldsymbol{\tau}^{k+1} = \boldsymbol{\tau}^k$. We first derive upper bounds on the last two terms in (2.42) for these two cases seperately. On the one hand, if $\mathrm{mod}\,(k+1, T) \neq 0$, we have that

$$\begin{aligned} &4L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^{k+1} - \mathbf{J}\boldsymbol{\tau}^{k+1}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^{k+1} - \mathbf{x}^*\right\|^2\right] \\ &= 4L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^{k} - \mathbf{J}\boldsymbol{\tau}^{k}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^{k} - \mathbf{x}^*\right\|^2\right]. \end{aligned} \tag{2.43}$$

On the other hand, if $\mathrm{mod}(k+1, T) = 0$, we have that

$$\begin{aligned} &4L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^{k+1} - \mathbf{J}\boldsymbol{\tau}^{k+1}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^{k+1} - \mathbf{x}^*\right\|^2\right] \\ &= 4L^2\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right]. \end{aligned} \tag{2.44}$$

Therefore, combining (2.43) and (2.44), we have that $\forall k \geq 0$:

$$\begin{aligned} &4L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^{k+1} - \mathbf{J}\boldsymbol{\tau}^{k+1}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^{k+1} - \mathbf{x}^*\right\|^2\right] \\ &\leq 4L^2\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right] \\ &\quad + 4L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^{k} - \mathbf{J}\boldsymbol{\tau}^{k}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^{k} - \mathbf{x}^*\right\|^2\right] \end{aligned} \tag{2.45}$$

Next, we apply (2.45) in (2.42) to obtain

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{v}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\right\|^2\right] &\leq 8L^2\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\right\|^2\right] + 8L^2\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right] \\ &\quad + 4L^2\mathbb{E}\left[\left\|\boldsymbol{\tau}^{k} - \mathbf{J}\boldsymbol{\tau}^{k}\right\|^2\right] + 4L^2\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^{k} - \mathbf{x}^*\right\|^2\right]. \end{aligned} \tag{2.46}$$

The proof follows by using (2.7), (2.10) and Lemma 2.5.11 in (2.46). □

2.5.4.2 Proof of Theorem 2.3.2

We now apply the upper bounds on the variance of the gradient estimator $\mathbf{v}^k$ in **GT-SVRG** obtained in the previous subsection to refine the inequalities derived for the general dynamical system (2.4)-(2.5) in Section 2.5 and establish the complexity for **GT-SVRG**. We first apply the upper bound on $\mathbb{E}[\|\mathbf{v}^k - \nabla \mathbf{f}(\mathbf{x}^k)\|^2]$ in Lemma 2.5.11 to (2.10) to obtain $\forall k \geq 0$:

$$\begin{aligned} \mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right] &\leq L^2\alpha\left(\frac{1}{\mu} + \frac{4\alpha}{n}\right)\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + \left(1 - \mu\alpha + \frac{4L^2}{n}\alpha^2\right)\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k} - \mathbf{x}^*\right\|^2\right] \\ &\quad + \frac{4L^2\alpha^2}{n}\mathbb{E}\left[\left\|\boldsymbol{\tau}^{k} - \mathbf{J}\boldsymbol{\tau}^{k}\right\|^2\right] + \frac{4L^2\alpha^2}{n}\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^{k} - \mathbf{x}^*\right\|^2\right]. \end{aligned} \tag{2.47}$$

If $0 < \alpha \leq \frac{1}{4\mu}$, we have $(\frac{1}{\mu} + \frac{4\alpha}{n}) \leq \frac{2}{\mu}$; if $0 < \alpha \leq \frac{n\mu}{8L^2}$, we have $1 - \mu\alpha + \frac{4L^2}{n}\alpha^2 \leq 1 - \frac{\mu\alpha}{2}$. Therefore, if $0 < \alpha \leq \frac{\mu}{8L^2}$, we have $k \geq 0$:

$$\begin{aligned}\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}^*\right\|^2\right] \leq \frac{2L^2\alpha}{\mu}\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + \left(1 - \frac{\mu\alpha}{2}\right)\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] \\ + \frac{4L^2\alpha^2}{n}\mathbb{E}\left[\left\|\boldsymbol{\tau}^k - \mathbf{J}\boldsymbol{\tau}^k\right\|^2\right] + \frac{4L^2\alpha^2}{n}\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^k - \mathbf{x}^*\right\|^2\right]. \qquad (2.48)\end{aligned}$$

Next, we apply the upper bounds on $\mathbb{E}[\|\mathbf{v}^k - \nabla\mathbf{f}(\mathbf{x}^k)\|^2]$ and $\mathbb{E}[\|\mathbf{v}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2]$ in Lemma 2.5.11 and Corollary 2.5.2 to Lemma 2.5.6 and obtain: $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\right\|^2\right] \leq \frac{120L^2}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] + \frac{87L^2}{1-\lambda^2}\mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] \\ + \frac{3+\lambda^2}{4}\mathbb{E}\left[\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right] \\ + \frac{38L^2}{1-\lambda^2}\mathbb{E}\left[\left\|\boldsymbol{\tau}^k - \mathbf{J}\boldsymbol{\tau}^k\right\|^2\right] + \frac{38L^2}{1-\lambda^2}\mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^k - \mathbf{x}^*\right\|^2\right], \qquad (2.49)\end{aligned}$$

if $0 < \alpha \leq \frac{1-\lambda^2}{14\sqrt{2}L}$. Now, we write Lemma 2.6, (2.48) and (2.49) jointly in an entry-wise linear matrix inequality that characterizes the evolution of `GT-SVRG` in the following proposition.

Proposition 2.5.2. *Let Assumptions 2.3.1, 2.3.2 and 2.3.3 hold a nd Consider the iterates $\{\mathbf{x}^k\}, \{\mathbf{y}^k\}, \{\mathbf{v}^k\}$ generated by `GT-SVRG`. If the step-size α follows $0 < \alpha \leq \frac{\mu(1-\lambda^2)}{14\sqrt{2}L^2}$, then the following linear matrix inequality hold $\forall k \geq 0$:*

$$\mathbf{u}^{k+1} \leq \mathbf{R}_\alpha \mathbf{u}^k + \mathbf{H}_\alpha \widetilde{\mathbf{u}}^k, \qquad (2.50)$$

where $\mathbf{u}^k, \widetilde{\mathbf{u}}^k \in \mathbb{R}^3$ and $\mathbf{R}_\alpha, \mathbf{H}_\alpha \in \mathbb{R}^{3\times 3}$ are defined as

$$\mathbf{u}^k = \begin{bmatrix} \mathbb{E}\left[\left\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\right\|^2\right] \\ \mathbb{E}\left[n\left\|\overline{\mathbf{x}}^k - \mathbf{x}^*\right\|^2\right] \\ \mathbb{E}\left[L^{-2}\left\|\mathbf{y}^k - \mathbf{J}\mathbf{y}^k\right\|^2\right] \end{bmatrix}, \qquad \widetilde{\mathbf{u}}^k = \begin{bmatrix} \mathbb{E}\left[\left\|\boldsymbol{\tau}^k - \mathbf{J}\boldsymbol{\tau}^k\right\|^2\right] \\ \mathbb{E}\left[n\left\|\overline{\boldsymbol{\tau}}^k - \mathbf{x}^*\right\|^2\right] \\ 0 \end{bmatrix},$$

$$\mathbf{R}_\alpha = \begin{bmatrix} \frac{1+\lambda^2}{2} & 0 & \frac{2\alpha^2L^2}{1-\lambda^2} \\ \frac{2L^2\alpha}{\mu} & 1 - \frac{\mu\alpha}{2} & 0 \\ \frac{120}{1-\lambda^2} & \frac{87}{1-\lambda^2} & \frac{3+\lambda^2}{4} \end{bmatrix}, \qquad \mathbf{H}_\alpha = \begin{bmatrix} 0 & 0 & 0 \\ \frac{4L^2\alpha^2}{n} & \frac{4L^2\alpha^2}{n} & 0 \\ \frac{38}{1-\lambda^2} & \frac{38}{1-\lambda^2} & 0 \end{bmatrix}.$$

Note that T is the number of the inner loop iterations of `GT-SVRG`. We will show that the subsequence $\{\mathbf{u}^{tT}\}_{t\geq 0}$ of $\{\mathbf{u}^k\}_{k\geq 0}$, which corresponds to the outer loop updates of `GT-SVRG`, converges to zero linearly, based on which the total complexity of `GT-SVRG` will be established, in terms of the number of

total component gradient computations required at each node to find the solution $\mathbf{x}^*$. We now recall from Algorithm 2 that $\forall k \geq 0$, $\boldsymbol{\tau}^{k+1} = \mathbf{x}_i^{k+1}$ if $\text{mod}(k+1, T) = 0$; else $\boldsymbol{\tau}^{k+1} = \boldsymbol{\tau}^k$. Therefore, $\forall t \geq 0$ and $tT \leq k \leq (t+1)T - 1$, we have $\boldsymbol{\tau}^k = \mathbf{x}^{tT}$. Based on this discussion, (2.50) can be rewritten as the following dynamical system with delays:

$$\mathbf{u}^{k+1} \leq \mathbf{R}_\alpha \mathbf{u}^k + \mathbf{H}_\alpha \mathbf{u}^{tT}, \qquad \forall k \in [tT, (t+1)T - 1], \qquad \forall t \geq 0.$$

We recursively apply the above inequality over k to obtain the evolution of the outer loop iterations $\{\mathbf{u}^{tT}\}_{t\geq 0}$:

$$\mathbf{u}^{(t+1)T} \leq \left(\mathbf{R}_\alpha^T + \sum_{l=0}^{T-1} \mathbf{R}_\alpha^l \mathbf{H}_\alpha\right) \mathbf{u}^{tT}, \qquad \forall t \geq 0. \tag{2.51}$$

Clearly, to show the linear decay of $\{\mathbf{u}^{tT}\}_{t\geq 0}$, it sufficies to find the range of α such that $\rho\big(\mathbf{R}_\alpha^T + \sum_{l=0}^{T-1} \mathbf{R}_\alpha^l \mathbf{H}_\alpha\big) < 1$. To this aim, we first derive the range of α such that $\rho(\mathbf{R}_\alpha) < 1$.

Lemma 2.5.12. *Let Assumptions 2.3.1, 2.3.2, 2.3.3 hold and consider the system matrix $\mathbf{R}_\alpha$ defined in Proposition 2.5.2. If the step-size α follows $0 < \alpha \leq \frac{(1-\lambda^2)^2}{187QL}$, then*

$$\rho(\mathbf{R}_\alpha) \leq |||\, \mathbf{R}_\alpha \,|||_\infty^{\boldsymbol{\delta}} \leq 1 - \frac{\mu\alpha}{4}, \tag{2.52}$$

where $\boldsymbol{\delta} = \left[1, 8Q^2, \frac{6528Q^2}{(1-\lambda^2)^2}\right]^\top$.

Proof. In the light of Lemma 2.5.10, we solve for the range of α and a positive vector $\boldsymbol{\delta} = [\delta_1, \delta_2, \delta_3]$ such that the following entry-wise linear matrix inequality holds:

$$\mathbf{R}_\alpha \boldsymbol{\delta} \leq \left(1 - \frac{\mu\alpha}{4}\right) \boldsymbol{\delta},$$

which can be written equivalently as

$$\frac{\mu\alpha}{4} + \frac{2L^2\alpha^2}{1-\lambda^2}\frac{\delta_3}{\delta_1} \leq \frac{1-\lambda^2}{2}, \tag{2.53}$$

$$8Q^2\delta_1 \leq \delta_2, \tag{2.54}$$

$$\frac{\mu\alpha}{4} \leq \frac{1-\lambda^2}{4} - \frac{120}{1-\lambda^2}\frac{\delta_1}{\delta_3} - \frac{87}{1-\lambda^2}\frac{\delta_2}{\delta_3}. \tag{2.55}$$

Based on (2.54), we set $\delta_1 = 1$ and $\delta_2 = 6Q^2$. With δ_1 and δ_2 being fixed, we next choose $\delta_3 > 0$ such that the RHS of (2.55) is positive, i.e, $\frac{1-\lambda^2}{4\delta_3}\big(\delta_3 - \frac{480+2784Q^2}{(1-\lambda^2)^2}\big) > 0$. It suffices to set $\delta_3 = \frac{6528Q^2}{(1-\lambda^2)^2}$. Now, with the previously fixed values of $\delta_1, \delta_2, \delta_3$, in order to make (2.55) hold, it suffices to choose α such that $0 < \alpha \leq \frac{1-\lambda^2}{2\mu}$. Similary, it can be verified that in order to make (2.53) hold, it sufficies to make α satisfy $0 < \alpha \leq \frac{(1-\lambda^2)^2}{187QL}$, which completes the proof. □

We note that if the step-size α satisfies the condition in Lemma 2.5.12, we have $\rho(\mathbf{R}_\alpha) < 1$. Moreover, since $\mathbf{R}_\alpha$ is nonnegative, we have that

$$\sum_{l=0}^{T-1} \mathbf{R}_\alpha^l \leq \sum_{l=0}^{\infty} \mathbf{R}_\alpha^l = (\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}.$$

Therefore, the following from (2.51), we have:

$$\mathbf{u}^{(t+1)T} \leq \left(\mathbf{R}_\alpha^T + (\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha\right)\mathbf{u}^{tT}, \quad \forall t \geq 0. \tag{2.56}$$

The rest of the convergence analysis is to derive the condition on the the number of each inner iterations T and the step-size α of `GT-SVRG` such that the following inequality holds:

$$\rho(\mathbf{R}_\alpha^T + (\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha) < 1.$$

We first show that $(\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha$ is sufficiently small under an appropriate weighted matrix norm in the light of Lemma 2.5.10.

Lemma 2.5.13. *Let Assumptions 2.3.1, 2.3.2 and 2.3.3 hold. Consider the system matrices $\mathbf{R}_\alpha, \mathbf{H}_\alpha$ defined in Proposition 2.5.2. If the step-size α follows $0 < \alpha \leq \frac{(1-\lambda^2)^2}{187QL}$, then*

$$\left|\!\left|\!\left| (\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha \right|\!\right|\!\right|_\infty^{\mathbf{q}} \leq 0.66,$$

where $\mathbf{q} = \left[1, 1, \frac{1453}{(1-\lambda^2)^2}\right]^\top$.

Proof. We start by deriving an entry-wise upper bound for the matrix $(\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}$. Note that the determinant of $(\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}$ is given by

$$\det(\mathbf{I}_3 - \mathbf{R}_\alpha) = \frac{(1-\lambda^2)^2\mu\alpha}{16} - \frac{348L^4\alpha^3}{\mu(1-\lambda^2)^2} - \frac{120\alpha^3\mu L^2}{(1-\lambda^2)^2}.$$

It can be verified that if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{187QL}$,

$$\det(\mathbf{I}_3 - \mathbf{R}_\alpha) \geq \frac{(1-\lambda^2)^2\mu\alpha}{32}. \tag{2.57}$$

Then we derive an entry-wise upper bound for $\mathrm{adj}(\mathbf{I}_3 - \mathbf{R}_\alpha)$, where $\mathrm{adj}(\cdot)$ denotes the adjugate of the argument matrix and we denote $[\mathrm{adj}(\cdot)]_{i,j}$ as its i,jth entry:

$$\begin{aligned}
&[\mathrm{adj}(\mathbf{I}_3 - \mathbf{R}_\alpha)]_{1,2} = \frac{174L^2\alpha^2}{(1-\lambda^2)^2}, && [\mathrm{adj}(\mathbf{I}_3 - \mathbf{R}_\alpha)]_{1,3} = \frac{\mu L^2\alpha^3}{1-\lambda^2},\\
&[\mathrm{adj}(\mathbf{I}_3 - \mathbf{R}_\alpha)]_{2,2} \leq \frac{(1-\lambda^2)^2}{8}, && [\mathrm{adj}(\mathbf{I}_3 - \mathbf{R}_\alpha)]_{2,3} = \frac{4L^4\alpha^3}{\mu(1-\lambda^2)},\\
&[\mathrm{adj}(\mathbf{I}_3 - \mathbf{R}_\alpha)]_{3,2} = \frac{87}{2}, && [\mathrm{adj}(\mathbf{I}_3 - \mathbf{R}_\alpha)]_{3,3} = \frac{\mu\alpha(1-\lambda^2)}{4}.
\end{aligned}$$

With the help of the above calculations, an entry-wise upper bound for $(\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha = \frac{\mathrm{adj}(\mathbf{I}_3-\mathbf{R}_\alpha)}{\det(\mathbf{I}_3-\mathbf{R}_\alpha)}\mathbf{H}_\alpha$ can be obtained, i.e., if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{187QL}$, we have

$$(\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha \leq \begin{bmatrix} 0.039 & 0.039 & 0 \\ 0.23 & 0.23 & 0 \\ \frac{334}{(1-\lambda^2)^2} & \frac{334}{(1-\lambda^2)^2} & 0 \end{bmatrix}.$$

Using Lemma 2.5.10 in a similar way as the proof of Lemma 2.5.12, it can be verified that $\big((\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha\big)\mathbf{q} \leq 0.66\mathbf{q}$, where $\mathbf{q} = [1, 1, \frac{1453}{(1-\lambda^2)^2}]^\top$, which completes the proof. □

Note that we use two different weighted matrix norms to bound $\mathbf{R}_\alpha$ and $(\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha$ respectively in Lemma 2.5.12 and 2.5.13, i.e., $|||\cdot|||_\infty^{\boldsymbol{\delta}}$ and $|||\cdot|||_\infty^{\mathbf{q}}$, where $\boldsymbol{\delta} = [1, 8Q^2, \frac{6528Q^2}{(1-\lambda^2)^2}]^\top$ and $\mathbf{q} = [1, 1, \frac{1453}{(1-\lambda^2)^2}]^\top$. It can be verified that [36]: $\forall \mathbf{X} \in \mathbb{R}^{3\times 3}$,

$$|||\mathbf{X}|||_\infty^{\mathbf{q}} \leq 8Q^2 |||\mathbf{X}|||_\infty^{\boldsymbol{\delta}}. \tag{2.58}$$

We next show the linear convergence of the outer loop of **GT-SVRG**, i.e., the linear decay of the subsequence $\{\mathbf{u}^{tT}\}_{t\geq 0}$ of $\{\mathbf{u}^k\}_{k\geq 0}$, where T is the number of inner loop iterations.

Proof of Theorem 2.3.2. Consider the iterates $\{\mathbf{u}^k\}$ generated by **GT-SVRG** (defined in Proposition 2.5.2) and recall the recursion in (2.56): $\forall t \geq 0, \mathbf{u}^{(t+1)T} \leq \big(\mathbf{R}_\alpha^T + (\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha\big)\mathbf{u}^{tT}$. Note that the weighted vector norm $\|\cdot\|_\infty^{\mathbf{q}}$ induces the weighted matrix norm $|||\cdot|||_\infty^{\mathbf{q}}$ [36]. Then using Lemma 2.5.12, 2.5.13 and (2.58), If the step-size $\alpha = \frac{(1-\lambda^2)^2}{187QL}$ and the number of inner loop iterations $T = \frac{1496Q^2}{(1-\lambda^2)^2}\log(200Q)$, then we have: $\forall t \geq 0$,

$$\begin{aligned} \left\|\mathbf{u}^{(t+1)T}\right\|_\infty^{\mathbf{q}} &\leq \left|\left|\left| \mathbf{R}_\alpha^T + (\mathbf{I}_3 - \mathbf{R}_\alpha)^{-1}\mathbf{H}_\alpha \right|\right|\right|_\infty^{\mathbf{q}} \left\|\mathbf{u}^{tT}\right\|_\infty^{\mathbf{q}} \\ &\leq \left(|||\mathbf{R}_\alpha^T|||_\infty^{\mathbf{q}} + 0.66\right)\left\|\mathbf{u}^{tT}\right\|_\infty^{\mathbf{q}} \\ &\leq \left(8Q^2\big(|||\mathbf{R}_\alpha|||_\infty^{\boldsymbol{\delta}}\big)^T + 0.66\right)\left\|\mathbf{u}^{tT}\right\|_\infty^{\mathbf{q}} \\ &\leq 0.7\left\|\mathbf{u}^{tT}\right\|_\infty^{\mathbf{q}}, \end{aligned} \tag{2.59}$$

Clearly, (2.59) shows that the outer loop of **GT-SVRG**, i.e., $\{\mathbf{x}^{tT}\}_{t\geq 0}$, converges to an ϵ-optimal solution with $\mathcal{O}(\log\frac{1}{\epsilon})$ iterations. We further note that in each inner loop of **GT-SVRG**, each node i computes (m_i+2T) local component gradients. Therefore, the total number of component gradient computations at each node required is $\mathcal{O}\big(\big(M + \frac{Q^2\log Q}{(1-\lambda)^2}\big)\log\frac{1}{\epsilon}\big)$, where M is the largest number of data points over all nodes and the proof is complete. □

2.6 Conclusion

In this chapter, we present the **GT-VR** framework that allows flexible and appropriate constructions of decentralized stochastic variance-reduced gradient methods over weight-balanced directed graphs with the help of gradient tracking techniques. For definiteness, we develop proper decentralized versions of the centralized SAGA and SVRG algorithms, namely **GT-SAGA** and **GT-SVRG**. It is shown that they achieve linear convergence to the optimal solution for smooth and strongly convex problems. Furthermore, we show that **GT-SAGA** and **GT-SVRG** in the big data regimes achieve non-asymptotic topology-independent linear speedups compared with the centralized SAGA and SVRG that execute on a single node.

Chapter 3

Decentralized Smooth Non-Convex Finite-Sum Optimization

This chapter focuses on decentralized smooth non-convex empirical risk minimization problems. In particular, we consider n nodes communicating over a balanced directed graph, where each node i has access to a local, private, collection of m smooth component functions $\{f_{i,j} : \mathbb{R}^p \to \mathbb{R}\}_{j=1}^m$ that are possibly *non-convex*. Each $f_{i,j}$ can be viewed as a cost incurred by the j-th data sample at the i-th node. Our goal here is to have the networked nodes agree on *a first-order stationary point* of the average of all component functions across the nodes via local computation and decentralized communication. Under the `GT-VR` framework developed in Chapter 2, we propose and analyze two decentralized stochastic variance-reduced algorithms, `GT-SARAH` and `GT-SAGA`, to tackle this smooth non-convex finite-sum formulation. Specifically, we show that the gradient complexity of `GT-SARAH` matches that of the centralized optimal methods for this problem class in big-data regimes like data centers, while `GT-SAGA` exhibits superior performance compared with `GT-SARAH` and other existing approaches in large-scale network regimes like Internet of Things (IoT).

3.1 Introduction

We consider decentralized finite-sum minimization of $N := nm$ cost functions that takes the following form:

$$\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \qquad f_i(\mathbf{x}) := \frac{1}{m}\sum_{j=1}^{m} f_{i,j}(\mathbf{x}), \tag{3.1}$$

where each $f_i : \mathbb{R}^p \to \mathbb{R}$, further decomposed as the average of m component costs $\{f_{i,j}\}_{j=1}^m$, is available only at the i-th node in a network of n nodes. The network is abstracted as a directed graph $\mathcal{G} := \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} := \{1, \cdots, n\}$ is the set of node indices and $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$ is the collection of ordered pairs $(i, r), i, r \in \mathcal{V}$, such that node r sends information to node i. We adopt the convention that $(i, i) \in \mathcal{E}, \forall i \in \mathcal{V}$. Each node

in the network is restricted to local computation and communication with its neighbors. Throughout the chapter, we focus on the case where each $f_{i,j}$ is differentiable, not necessarily convex, and F is bounded below. This formulation often appears in decentralized empirical risk minimization, where each local cost f_i can be considered as an empirical risk computed over a finite number of m local data samples [31], and lies at the heart of many modern machine learning problems [7, 27]. Examples include non-convex linear models and neural networks. When the local data size m is large, evaluating the exact gradient ∇f_i of each local cost at each iteration becomes computationally expensive and methods that efficiently sample each local data batch are preferable. We are thus interested in designing fast stochastic gradient algorithms to find an ϵ-accurate first-order stationary point $\widehat{\mathbf{x}} \in \mathbb{R}^p$ such that $\mathbb{E}\big[\|\nabla F(\widehat{\mathbf{x}})\|^2\big] \leq \epsilon^2$.

Towards Problem (3.1), `DSGD` [19, 37–39], a decentralized version of stochastic gradient descent (`SGD`) [7, 47, 128], is often used to address the large-scale and decentralized nature of the data. `DSGD` is popular for several inference and learning tasks due to its simplicity of implementation and speedup in comparison to its centralized counterparts [2]. `DSGD` and its variants have been been extensively studied for different computation and communication needs, e.g., momentum [40], directed graphs [41], escaping saddle-points [42, 43], zeroth-order schemes [44], swarming-based implementations [45], and constrained problems [46]. The performance of `DSGD` for the non-convex Problem (3.1) however suffers from three major challenges: (i) the non-degenerate variance of the stochastic gradients at each node; (ii) the dissimilarity among the local functions across the nodes; and (iii) the transient time to reach the network topology independent region. To elaborate these issues, we recap `DSGD` for Problem (3.1) and its convergence results as follows. Let $\mathbf{x}_i^k \in \mathbb{R}^p$ denote the iterate of `DSGD` at node i and iteration k. At each node i, `DSGD` performs [37, 39]

$$\mathbf{x}_i^{k+1} = \sum_{r=1}^{n} \underline{w}_{ir} \mathbf{x}_r^k - \alpha \cdot \mathbf{g}_i^k, \qquad k \geq 0, \tag{3.2}$$

where $\underline{\mathbf{W}} = \{\underline{w}_{ir}\} \in \mathbb{R}^{n \times n}$ is a weight matrix that respects the network topology, while $\mathbf{g}_i^k \in \mathbb{R}^p$ is a stochastic gradient such that $\mathbb{E}[\mathbf{g}_i^k | \mathbf{x}_i^k] = \nabla f_i(\mathbf{x}_i^k)$. Assuming the *bounded variance* of each local stochastic gradient $\mathbf{g}_i^k$, the *bounded dissimilarity* between the local and the global gradient [2], i.e., for some $\nu > 0$ and $\zeta > 0$,

$$\sup_{i \in \mathcal{V}, k \geq 0} \mathbb{E}\Big[\big\|\mathbf{g}_i^k - \nabla f_i(\mathbf{x}_i^k)\big\|^2\Big] \leq \nu^2 \qquad \text{and} \qquad \sup_{\mathbf{x} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq \zeta^2, \tag{3.3}$$

and L-smoothness of each f_i, it is shown in [2] that, for small enough $\alpha > 0$,

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\Big[\big\|\nabla F(\overline{\mathbf{x}}^k)\big\|^2\Big] = \mathcal{O}\left(\frac{F(\overline{\mathbf{x}}^0) - F^*}{\alpha K} + \frac{\alpha L \nu^2}{n} + \frac{\alpha^2 L^2 \nu^2}{1-\lambda} + \frac{\alpha^2 L^2 \zeta^2}{(1-\lambda)^2}\right), \tag{3.4}$$

where $\overline{\mathbf{x}}^k := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^k$ and $(1-\lambda) \in (0, 1]$ is the spectral gap of the weight matrix $\underline{\mathbf{W}}$. It then follows that [2] for K *large enough* and with an appropriate step-size α, `DSGD` finds an ϵ-accurate first-order stationary point

of F in $\mathcal{O}(\nu^2 L\epsilon^{-4})$ stochastic gradient computations across all nodes and therefore achieves *asymptotic* linear speedup compared to the centralized `SGD` [7,47] that executes at a single node. Clearly, there are three issues with the convergence properties of `DSGD`:

- Due to the non-degenerate stochastic gradient variance, the gradient complexity of `DSGD` does not match that of the centralized near-optimal variance-reduced methods when minimizing a finite-sum of smooth non-convex functions [48–50].
- The bounded dissimilarity assumption on the local and global gradients [2,41,43] or the coercivity of each local function [42] is essential for establishing the convergence of `DSGD`. In fact, a counterexample has been shown in [15] that *`DSGD` diverges for any constant step-size* when these types of assumptions are violated. Furthermore, the practical performance of `DSGD` degrades significantly when the local and the global gradients are substantially different, i.e., when the data distributions across the nodes are largely heterogeneous [3,4,20].
- `DSGD` achieves linear speedup only *asymptotically*, i.e., after a finite number of transient iterations that is a polynomial function of n, ν, ζ, L, and $(1-\lambda)$ [2,40,116].

In this chapter, we show that the `GT-VR` framework presented in Chapter 2 provably addresses the aforementioned challenges posed by `DSGD` in decentralized non-convex finite-sum optimization. We further discuss trade-offs between different instantiations of `GT-VR` in this context.

3.2 Stochastic recursive variance reduction

In this section, we present and analyze the convergence properties of an instance of the `GT-VR` framework that uses a recursive variance reduction technique [58].

3.2.1 Main contributions

We propose `GT-SARAH`, a novel decentralized stochastic variance-reduced gradient method that provably addresses the aforementioned challenges posed by `DSGD`. `GT-SARAH` is based on a *local `SARAH`-type gradient estimator* [48,49], which removes the variance incurred by the local stochastic gradients, and *global gradient tracking (`GT`)* [55,65,129], that fuses the gradient estimators across the nodes such that the bounded dissimilarity or the coercivity type assumptions are not required. Our main technical contributions for `GT-SARAH` are summarized in the following.

- We show that `GT-SARAH`, under appropriate algorithmic parameters, finds an ϵ-accurate first-order stationary point $\widehat{\mathbf{x}}$ of F such that $\mathbb{E}\big[\|\nabla F(\widehat{\mathbf{x}})\|^2\big] \leq \epsilon^2$ in at most

$$\mathcal{H}_R := \mathcal{O}\big(\max\big\{N^{1/2}, n(1-\lambda)^{-2}, n^{2/3}m^{1/3}(1-\lambda)^{-1}\big\}L\epsilon^{-2}\big)$$

  component gradient computations across all nodes. The gradient complexity $\mathcal{H}_R$ significantly outperforms that of the existing decentralized stochastic gradient algorithms for Problem (3.1); see Table 3.1.

- In a big-data regime such that $n = \mathcal{O}(N^{1/2}(1-\lambda)^3)$, the gradient complexity $\mathcal{H}_R$ of `GT-SARAH` reduces to $\widetilde{\mathcal{H}}_R := \mathcal{O}(N^{1/2}L\epsilon^{-2})$. We emphasize that $\widetilde{\mathcal{H}}_R$ is independent of the network topology and matches that of the centralized near-optimal variance-reduced methods [48–50] under a slightly stronger smoothness assumption; see Remark 3.2.1 for details. Furthermore, since `GT-SARAH` computes n gradients in parallel at each iteration, its per-node gradient complexity in this regime is $\mathcal{O}(N^{1/2}n^{-1}\epsilon^{-2})$, demonstrating a *non-asymptotic linear speedup* compared with the aforementioned centralized near-optimal methods [48–50] that perform all gradient computations at a single node. To the best of our knowledge, `GT-SARAH` is the first decentralized method that achieves this property for Problem (3.1).

- We show that choosing the local minibatch size of `GT-SARAH` judiciously balances the trade-offs between the gradient and communication complexity; see Corollary 3.2.1 and Subsection 3.2.4.3 for details.

- We establish that all nodes in `GT-SARAH` asymptotically achieve consensus and converge to a first-order stationary point of F over infinite time horizon in the almost sure and mean-squared sense.

3.2.2 Related work

Several algorithms have been proposed to improve certain aspects of `DSGD`. For example, a stochastic variant of `EXTRA` [68], `Exact Diffusion` [20], and `NIDS` [69], called `D2` [3], removes the bounded dissimilarity assumption in `DSGD` based on a bias-correction principle. `DSGT` [4], introduced in [67] for smooth and strongly convex problems, achieves a similar theoretical performance as `D2` via gradient tracking [54–56, 72], but with more general choices of weight matrices. Reference [130] establishes asymptotic properties of a decentralized stochastic primal-dual algorithm for smooth convex problems. Reference [131] develops decentralized primal-dual communication sliding algorithms that achieve communication efficiency for convex and possibly nonsmooth problems. These methods however are subject to the non-degenerate variance of the stochastic gradients. Inspired by the variance-reduction techniques for centralized stochastic optimization [48–50, 57, 58, 62–64, 132–134], decentralized variance-reduced methods for smooth and strongly-convex problems have been proposed recently, e.g., in [22–24, 31, 120]; in particular, the integration of gradient tracking and variance reduction described here was introduced in [24, 31] to obtain linear convergence.

Table 3.1: A comparison of the gradient complexities of the-state-of-the-art decentralized stochastic gradient methods to minimize a sum of $N = nm$ smooth non-convex functions equally divided among n nodes. The gradient complexity is in terms of the total number of component gradient computations across all nodes to find a first-order stationary point $\widehat{\mathbf{x}} \in \mathbb{R}^p$ such that $\mathbb{E}\big[\|\nabla F(\widehat{\mathbf{x}})\|^2\big] \leq \epsilon^2$. In the table, ν^2 denotes the bounded variance of the stochastic gradients described in (3.3), $(1-\lambda) \in (0,1]$ is the spectral gap of the network weight matrix and L is the smoothness parameter of the cost functions. We note that the complexities of `DSGD`, `D2`, `DSGT` in the table are established in the setting of stochastic first-order oracles, which is more general than the finite-sum formulation considered here. Moreover, the complexities of `DSGD`, `D2`, `DSGT` in the table are stated in the regime that ϵ is small enough for simplicity; see [2–4] for their precise expressions. Finally, we note that only the best possible gradient complexity of `GT-SARAH`, in the sense of Theorem 3.2.3, is presented in the table for conciseness; see Corollary 3.2.1 and Subsection 3.2.4.3 for detailed discussion on balancing the trade-offs between the gradient and communication complexity of `GT-SARAH`.

Algorithm	**Gradient complexity**	**Remarks**
`DSGD` [2]	$\mathcal{O}\left(\frac{\nu^2 L}{\epsilon^4}\right)$	bounded variance, bounded dissimilarity
`D2` [3]	$\mathcal{O}\left(\frac{\nu^2 L}{\epsilon^4}\right)$	bounded variance
`GT-DSGD` [4]	$\mathcal{O}\left(\frac{\nu^2 L}{\epsilon^4}\right)$	bounded variance
`D-GET` [135]	$\mathcal{O}\left(\frac{n^{1/2} N^{1/2} L^b}{(1-\lambda)^a \epsilon^2}\right)$	$a, b \in \mathbb{R}^+$ are not explicitly shown in [135]
`GT-SARAH` **(this work)**	$\mathcal{O}\left(\max\left\{N^{1/2}, \frac{n}{(1-\lambda)^2}, \frac{n^{2/3} m^{1/3}}{1-\lambda}\right\} \frac{L}{\epsilon^2}\right)$	See Theorem 3.2.3 and Corollary 3.2.1

A recent paper [135] proposes `D-GET` for Problem (3.1), which also considers local `SARAH`-type variance reduction and gradient tracking. In the following, we compare our work to [135] from a few major technical aspects.[1] *First*, the gradient complexity $\mathcal{H}_R$ of `GT-SARAH` improves that of `D-GET` in terms of the dependence on n and m; see Table 3.1. In particular, in a big-data regime, $n = \mathcal{O}(N^{1/2}(1-\lambda)^3)$, $\mathcal{H}_R$ matches the gradient complexity of the centralized near-optimal methods [48–50]; in contrast, the gradient complexity of `D-GET` is worse than that of the centralized near-optimal methods by a factor of $n^{1/2}$ even if the network is fully-connected. *Second*, the complexity results of `D-GET` are attained with a specific local minibatch size $m^{1/2}$. Conversely, we establish general complexity bounds of `GT-SARAH` with arbitrary local minibatch size and characterize the computation-communication trade-offs induced by different choices of the minibatch size. *Third*, the Lyapunov function based convergence analysis of `D-GET` does not show explicit dependence of several important problem parameters, such as $(1-\lambda)$ and L, while the analysis in this work reveals explicitly the dependence of all problem related parameters and sheds light on their implications. *Fourth*, we note that both `GT-SARAH` and `D-GET` achieve a worst case communication complexity of the form $\mathcal{O}((1-\lambda)^{-a} L^b \epsilon^{-2})$, independent of m and n, for some $a, b \in \mathbb{R}^+$. Since the dependence of a and b in `D-GET` are not explicit, it is unclear which algorithm achieves a lower communication complexity. *Finally*, [135] presents a variant of `D-GET` that is applicable to a more general online setting such as expected risk minimization.

[1] Note that [135] uses $\mathbb{E}[\|\nabla F(\widehat{\mathbf{x}})\|^2] \leq \epsilon$ as the performance metric, while we use $\mathbb{E}[\|\nabla F(\widehat{\mathbf{x}})\|^2] \leq \epsilon^2$ here. We state the complexities of `D-GET` established in [135] under our metric for consistency.

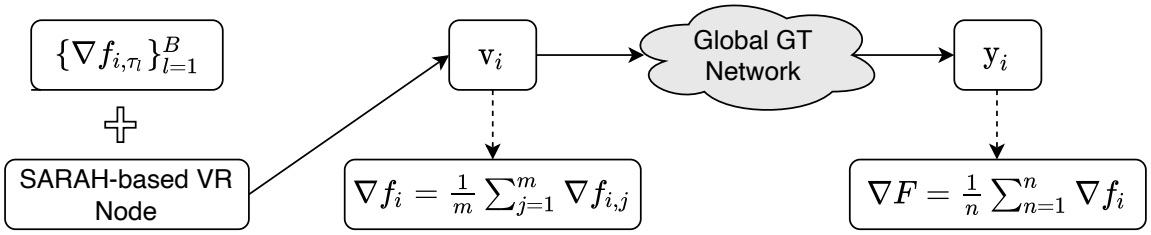

Figure 3.1: Each node i samples a minibatch of stochastic gradients $\{\nabla f_{i,\tau_l}\}_{l=1}^{B}$ at each iteration from its local data batch and computes an estimator $\mathbf{v}_i$ of its local batch gradient ∇f_i via a SARAH-type variance reduction (VR) procedure. These local gradient estimators $\mathbf{v}_i$'s are then fused over the network via a gradient tracking technique to obtain $\mathbf{y}_i$'s that approximate the global gradient ∇F.

3.2.3 The GT-SARAH algorithm

We now systematically build the proposed algorithm GT-SARAH and provide the basic intuition. We recall that the performance (3.4) of DSGD, in addition to the first term which is similar to that of the centralized batch gradient descent, has three additional bias terms. The second and third bias terms in (3.4) depend on the variance ν^2 of local stochastic gradients. A variance-reduced gradient estimation procedure of SARAH-type [48, 49], employed locally at each node i in GT-SARAH, removes ν^2. The last bias term in (3.4) is due to the dissimilarity ζ^2 between the local gradients $\{\nabla f_i\}_{i=1}^{n}$ and the global gradient ∇F. A dynamic fusion mechanism, called gradient tracking [54–56, 65, 73], removes ζ^2 by tracking the average of the local gradient estimators in GT-SARAH to learn the global gradient at each node. This process is illustrated in Fig. 3.1.

The complete implementation of GT-SARAH is summarized in Algorithm 3, where we assume that all nodes start from the same point $\overline{\mathbf{x}}^{0,1} \in \mathbb{R}^p$. GT-SARAH can be interpreted as a double loop method with an outer loop, indexed by s, and an inner loop, indexed by t. At the beginning of each outer loop s, GT-SARAH computes the local batch gradient $\mathbf{v}_i^{0,s} := \nabla f_i(\mathbf{x}_i^{0,s})$ at each node i. These batch gradients are then used to compute the first iteration of the global gradient tracker $\mathbf{y}_i^{1,s}$ and the state update $\mathbf{x}_i^{1,s}$. The three quantities, $\mathbf{v}_i^{0,s}, \mathbf{y}_i^{1,s}, \mathbf{x}_i^{1,s}$, set up the subsequent inner loop iterations. At each inner loop iteration $t \geq 1$, each node i samples two minibatch stochastic gradients from its local data that are used to construct the gradient estimator $\mathbf{v}_i^{t,s}$. We note that the gradient estimator is of recursive nature, i.e., it depends on $\mathbf{v}_i^{t-1,s}$ and the minibatch stochastic gradients evaluated at the current and the past states $\mathbf{x}_i^{t,s}$ and $\mathbf{x}_i^{t-1,s}$. The next step is to update $\mathbf{y}_i^{t+1,s}$ based on the gradient tracking protocol. Finally, the state $\mathbf{x}_i^{t+1,s}$ at each node i is computed as a convex combination of the states of the neighboring nodes followed by a descent in the direction of the gradient tracker $\mathbf{y}_i^{t+1,s}$. The latest updates $\mathbf{x}_i^{q+1,s}$, $\mathbf{y}_i^{q+1,s}$ and $\mathbf{v}_i^{q,s}$ then set up the next inner-outer loop cycle of GT-SARAH.

Algorithm 3 `GT-SARAH` at each node i

Require: $\mathbf{x}_i^{0,1} = \overline{\mathbf{x}}^{0,1} \in \mathbb{R}^p$, $\alpha \in \mathbb{R}^+$, $q \in \mathbb{Z}^+$, $S \in \mathbb{Z}^+$, $B \in \mathbb{Z}^+$, $\{\underline{w}_{ir}\}_{r=1}^n$, $\mathbf{y}_i^{0,1} = \mathbf{0}_p$, $\mathbf{v}_i^{-1,1} = \mathbf{0}_p$.

for $s = 1, 2, \cdots, S$ **do**

  $\mathbf{v}_i^{0,s} = \nabla f_i(\mathbf{x}_i^{0,s}) = \frac{1}{m}\sum_{j=1}^m \nabla f_{i,j}(\mathbf{x}_i^{0,s})$; ▷ batch gradient computation

  $\mathbf{y}_i^{1,s} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{y}_i^{0,s} + \mathbf{v}_i^{0,s} - \mathbf{v}_i^{-1,s}$ ▷ gradient tracking

  $\mathbf{x}_i^{1,s} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{x}_r^{0,s} - \alpha\mathbf{y}_i^{1,s}$ ▷ state update

  for $t = 1, 2, \cdots, q$ **do**

    for l in $\{1, \cdots, B\}$, choose $\tau_{i,l}^{t,s}$ uniformly at random from $\{1, \cdots, m\}$; ▷ sampling

    $\mathbf{v}_i^{t,s} = \frac{1}{B}\sum_{l=1}^B \left(\nabla f_{i,\tau_{i,l}^{t,s}}(\mathbf{x}_i^{t,s}) - \nabla f_{i,\tau_{i,l}^{t,s}}(\mathbf{x}_i^{t-1,s})\right) + \mathbf{v}_i^{t-1,s}$; ▷ SARAH

    $\mathbf{y}_i^{t+1,s} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{y}_r^{t,s} + \mathbf{v}_i^{t,s} - \mathbf{v}_i^{t-1,s}$; ▷ gradient tracking

    $\mathbf{x}_i^{t+1,s} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{x}_r^{t,s} - \alpha\mathbf{y}_i^{t+1,s}$; ▷ state update

  end for

  Set $\mathbf{x}_i^{0,s+1} = \mathbf{x}_i^{q+1,s}$; $\mathbf{y}_i^{0,s+1} = \mathbf{y}_i^{q+1,s}$; $\mathbf{v}_i^{-1,s+1} = \mathbf{v}_i^{q,s}$. ▷ next cycle

end for

3.2.4 Main convergence results

In this section, we present the main convergence results of `GT-SARAH` and discuss their implications. We make the following assumptions to establish the convergence properties of `GT-SARAH`.

Assumption 3.2.1. *Each local component cost $f_{i,j}$ is differentiable and $\{f_{i,j}\}_{j=1}^m$ satisfies a mean-squared smoothness property, i.e., for some $L > 0$,*

$$\frac{1}{m}\sum_{j=1}^m \|\nabla f_{i,j}(\mathbf{x}) - \nabla f_{i,j}(\mathbf{y})\|^2 \leq L^2\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall i \in \mathcal{V}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p. \tag{3.5}$$

In addition, the global cost F is bounded below, i.e., $F^* := \inf_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) > -\infty$.

It is clear that under Assumption 3.2.1, each f_i and F are L-smooth. We note that Assumption 3.2.1 is weaker than requiring each $f_{i,j}$ to be L-smooth.

Remark 3.2.1. The local mean-squared smoothness assumption (3.5), which is also used in the existing work [135], is slightly stronger than the smoothness assumption required by the existing lower bound $\Omega(N^{1/2}L\epsilon^{-2})$ [49, 136] and the centralized near-optimal methods [48–50] for finite-sum problems in the following sense. If we view Problem (3.1) as a centralized optimization problem, that is, all $f_{i,j}$'s are available at a single node, then the aforementioned lower bound and the convergence of the centralized near-optimal methods are established under the following assumption:

$$\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{i,j}(\mathbf{x}) - \nabla f_{i,j}(\mathbf{y})\|^2 \leq L^2\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p. \tag{3.6}$$

Clearly, (3.6) is implied by (3.5) but not vice versa. Due to this subtle difference, it is unclear whether the existing lower bound $\Omega(N^{1/2}L\epsilon^{-2})$ [49, 136] established under (3.6) remains valid under (3.5). Finally, we note that a lower bound result for decentralized deterministic first-order algorithms in the case of $m = 1$ can be found in [87].

Assumption 3.2.2. *The family* $\{\tau_{i,l}^{t,s} : t \in [1, q], s \geq 1, i \in \mathcal{V}, l \in [1, B]\}$ *of random variables is independent.*

Assumption 3.2.2 is standard in the stochastic optimization literature, e.g., [7, 49].

Assumption 3.2.3. *The nonnegative weight matrix* $\underline{\mathbf{W}} := \{\underline{w}_{ir}\} \in \mathbb{R}^{n\times n}$ *associated with the network* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *has positive diagonals and is primitive. Moreover,* $\underline{\mathbf{W}}$ *is doubly stochastic, i.e.,* $\underline{\mathbf{W}}\mathbf{1}_n = \mathbf{1}_n$ *and* $\mathbf{1}_n^\top \underline{\mathbf{W}} = \mathbf{1}_n^\top$.

An important consequence of Assumption 3.2.3 is that [56]

$$\lambda := \left\|\underline{\mathbf{W}} - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right\| = \lambda_2(\underline{\mathbf{W}}) \in [0, 1), \tag{3.7}$$

where $\lambda_2(\underline{\mathbf{W}})$ denotes the second largest singular value of $\underline{\mathbf{W}}$.[2] We term $(1 - \lambda)$ as the spectral gap of $\underline{\mathbf{W}}$ that characterizes the connectivity of the network [27].

Remark 3.2.2. Weight matrices satisfying Assumption 3.2.3 may be designed for the family of strongly-connected directed graphs that admit doubly-stochastic weights: (i) towards the primitivity requirement in Assumption 3.2.3, we note that if a graph is strongly-connected, then its associated weight matrix $\underline{\mathbf{W}}$ is irreducible [36, Theorem 6.2.14, 6.2.24] and $\underline{\mathbf{W}}$ is further primitive since it is nonnegative with positive diagonals [36, Lemma 8.5.4]; (ii) towards the doubly stochastic requirement in Assumption 3.2.3, we refer the readers to [125] for necessary and sufficient conditions under which a strongly connected directed graph admits doubly stochastic weights.

An important special case of this family is undirected connected graphs where doubly stochastic weights always exist and can be constructed in an efficient and decentralized manner, for instance, by the lazy Metroplis rule [27]. Hence, Assumption 3.2.3 is *more general* than the one required by `EXTRA`-based algorithms for decentralized optimization. For example, the weight matrix of `D2` needs to be symmetric and meet certain spectral properties [3] and is therefore not applicable to directed graphs.

We formally state the convergence results of `GT-SARAH` next, whose proofs are deferred to Subsection 3.2.6.2.

[2] We note that the relation in (3.7) may be established by following the definition of the spectral norm with the help of the primitivity and doubly stochasticity of $\underline{\mathbf{W}}$ and $\underline{\mathbf{W}}^\top\underline{\mathbf{W}}$, Perron-Frobenius theorem, and the spectral decomposition of $\underline{\mathbf{W}}^\top\underline{\mathbf{W}}$ [36, 56].

3.2.4.1 Asymptotic almost sure and mean-squared convergence

The following theorem shows the asymptotic convergence of `GT-SARAH`.

Theorem 3.2.1. *Let Assumptions 3.2.1-3.2.3 hold. Suppose that the step-size α, minibatch size B, and the inner-loop length q of* ***GT-SARAH*** *follow*

$$0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{4\sqrt{42}},\ \left(\frac{nB}{6q}\right)^{1/2},\ \left(\frac{4nB}{7nB+24q}\right)^{1/4}\frac{1-\lambda^2}{6}\right\}\frac{1}{2L},$$

where $B \in [1, m]$. Then we have: $\forall t \in [0, q]$, $\forall i \in \mathcal{V}$,

$$\mathbb{P}\left(\lim_{s\to\infty}\left\|\nabla F(\mathbf{x}_i^{t,s})\right\| = 0\right) = 1 \qquad \text{and} \qquad \lim_{s\to\infty}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_i^{t,s})\right\|^2\right] = 0,$$
$$\mathbb{P}\left(\lim_{s\to\infty}\left\|\mathbf{x}_i^{t,s} - \overline{\mathbf{x}}^{t,s}\right\| = 0\right) = 1 \qquad \text{and} \qquad \lim_{s\to\infty}\mathbb{E}\left[\left\|\mathbf{x}_i^{t,s} - \overline{\mathbf{x}}^{t,s}\right\|^2\right] = 0,$$

where $\overline{\mathbf{x}}^{t,s} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{t,s}$.

In addition to the mean-squared convergence that is standard in the stochastic optimization literature, the almost sure convergence in Theorem 3.2.1 guarantees that all nodes in `GT-SARAH` asymptotically achieve consensus and converge to a first-order stationary point of F on almost every sample path.

3.2.4.2 Complexities of `GT-SARAH` for finding first-order stationary points

We measure the outer-loop complexity of `GT-SARAH` in the following sense.

Definition 3.2.1. *Consider the sequence of random vectors $\{\mathbf{x}_i^{t,s}\}$ generated by* ***GT-SARAH****, at each node i. We say that* ***GT-SARAH*** *finds an ϵ-accurate first-order stationary point of F in S outer-loop iterations if*

$$\frac{1}{S(q+1)}\sum_{s=1}^{S}\sum_{t=0}^{q}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_i^{t,s})\right\|^2 + L^2\left\|\mathbf{x}_i^{t,s} - \overline{\mathbf{x}}^{t,s}\right\|^2\right] \leq \epsilon^2. \tag{3.8}$$

This is a standard metric that is concerned with the minimum of the stationary gaps and consensus errors over iterations in the mean-squared sense at each node [2, 3, 48–50]. In particular, if (3.8) holds and the output $\widehat{\mathbf{x}}$ of `GT-SARAH` is chosen uniformly at random from the set $\{\mathbf{x}_i^{t,s} : 0 \leq t \leq q, 1 \leq s \leq S, i \in \mathcal{V}\}$, then we have $\mathbb{E}[\|\nabla F(\widehat{\mathbf{x}})\|^2] \leq \epsilon^2$. We first provide the outer-loop iteration complexity of `GT-SARAH`.

Theorem 3.2.2. *Let Assumptions 3.2.1-3.2.3 hold. Suppose that the step-size α, minibatch size B, and the inner-loop length q of* ***GT-SARAH*** *follow*

$$0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{4\sqrt{42}},\ \left(\frac{nB}{6q}\right)^{1/2},\ \left(\frac{4nB}{7nB+24q}\right)^{1/3}\frac{1-\lambda^2}{6}\right\}\frac{1}{2L},$$

where $B \in [1, m]$. Then the number of the outer-loop iterations S required by ***GT-SARAH*** *to find an ϵ-accurate stationary point of F is at most*

$$\frac{1}{(q+1)\alpha L\epsilon^2}\left(4L\left(F(\overline{\mathbf{x}}^{0,1}) - F^*\right) + \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(\overline{\mathbf{x}}^{0,1})\right\|^2\right).$$

Using Theorem 3.2.2, the gradient and communication complexities of GT-SARAH can be readily established.

Theorem 3.2.3. *Let Assumptions 3.2.1-3.2.3 hold. Suppose that the step-size α and the length q of the inner loop of GT-SARAH are chosen as*[3]

$$q = \mathcal{O}\Big(\frac{m}{B}\Big) \quad \text{and} \quad \alpha = \mathcal{O}\bigg(\min\Big\{(1-\lambda)^2,\ \frac{n^{1/2}B}{m^{1/2}},\ \frac{n^{1/3}B^{2/3}(1-\lambda)}{m^{1/3}}\Big\}\frac{1}{L}\bigg), \tag{3.9}$$

where $B \in [1, m]$. Then GT-SARAH finds an ϵ-accurate stationary point of F in

$$\mathcal{H}_B := \mathcal{O}\bigg(\max\Big\{\frac{nB}{(1-\lambda)^2},\ N^{1/2},\ \frac{m^{1/3}n^{2/3}B^{1/3}}{1-\lambda}\Big\}\frac{\Delta}{\epsilon^2}\bigg)$$

component gradient computations across all nodes and

$$\mathcal{K}_B := \mathcal{O}\bigg(\max\Big\{\frac{1}{(1-\lambda)^2},\ \frac{m^{1/2}}{n^{1/2}B},\ \frac{m^{1/3}}{n^{1/3}B^{2/3}(1-\lambda)}\Big\}\frac{\Delta}{\epsilon^2}\bigg)$$

rounds of communication, where $\Delta := L\left(F(\overline{\mathbf{x}}^{0,1}) - F^*\right) + \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\overline{\mathbf{x}}^{0,1})\|^2$.

Remark 3.2.3. Theorem 3.2.3 holds for an arbitrary minibatch size $B \in [1, m]$.

Remark 3.2.4. The gradient complexity *at each node* of GT-SARAH is $\mathcal{H}_B/n$.

In view of Theorem 3.2.3, as the minibatch size B increases, the gradient complexity $\mathcal{H}_B$ (resp. the communication complexity $\mathcal{K}_B$) of GT-SARAH is non-decreasing (resp. non-increasing). The following corollary may be obtained from Theorem 3.2.3 by standard algebraic manipulations and shows that choosing the minibatch size B appropriately leads to favorable computation and communication trade-offs.

Corollary 3.2.1. *Let Assumptions 3.2.1-3.2.3 hold. Suppose that the step-size α and the inner-loop length q of GT-SARAH are chosen according to (3.9). We have the following complexity results.*

(i) If $B \in [1, \lfloor R \rfloor]$, where $R := \max\{m^{1/2}n^{-1/2}(1-\lambda)^3, 1\}$, then GT-SARAH attains the best possible, in the sense of Theorem 3.2.3, gradient complexity

$$\mathcal{H}_R := \mathcal{O}\bigg(\max\Big\{\frac{n}{(1-\lambda)^2},\ N^{1/2},\ \frac{m^{1/3}n^{2/3}}{1-\lambda}\Big\}\frac{\Delta}{\epsilon^2}\bigg); \tag{3.10}$$

moreover, when $B = \lfloor R \rfloor$, the corresponding communication complexity of GT-SARAH is

$$\mathcal{K}_R := \mathcal{O}\bigg(\max\Big\{\frac{1}{(1-\lambda)^2}, \min\Big\{\frac{m^{1/2}}{n^{1/2}}, \frac{1}{(1-\lambda)^3}\Big\}, \min\Big\{\frac{m^{1/3}}{n^{1/3}(1-\lambda)}, \frac{1}{(1-\lambda)^3}\Big\}\Big\}\frac{\Delta}{\epsilon^2}\bigg). \tag{3.11}$$

(ii) If $B \in [\lceil C \rceil, m]$, where $C := \max\{m^{1/2}n^{-1/2}(1-\lambda)^{3/2}, 1\}$, then GT-SARAH attains the best possible, in the sense of Theorem 3.2.3, communication complexity

$$\mathcal{K}_C := \mathcal{O}\bigg(\frac{1}{(1-\lambda)^2}\frac{\Delta}{\epsilon^2}\bigg); \tag{3.12}$$

[3]The $\mathcal{O}$ notation only hides universal constants that are independent of problem parameters.

moreover, when $B = \lceil C \rceil$*, the corresponding gradient complexity of* `GT-SARAH` *is*

$$\mathcal{H}_C := \mathcal{O}\left(\max\left\{ \frac{n}{(1-\lambda)^2}, \frac{N^{1/2}}{(1-\lambda)^{1/2}}, \frac{m^{1/3}n^{2/3}}{1-\lambda} \right\} \frac{\Delta}{\epsilon^2} \right). \tag{3.13}$$

Comparing (3.10) (3.11) with (3.13) (3.12), we clearly have $\mathcal{H}_R \leq \mathcal{H}_C$ and $\mathcal{K}_R \geq \mathcal{K}_C$.

3.2.4.3 Two regimes of practical significance

We now discuss the implications of the complexity results in Corollary 3.2.1 and the corresponding computation-communication trade-offs in the following regimes of practical significance.

- **Big-data regime:** $n = \mathcal{O}(N^{1/2}(1-\lambda)^3)$. In this regime, typical to large-scale machine learning, i.e., the total number of data samples N is very large, it can be verified that $\mathcal{H}_R$ reduces to $\widetilde{\mathcal{H}}_R := \mathcal{O}(N^{1/2}\Delta\epsilon^{-2})$ and $\mathcal{K}_R$ reduces to $\widetilde{\mathcal{K}}_R := \mathcal{O}((1-\lambda)^{-3}\Delta\epsilon^{-2})$. It is worth noting that $\widetilde{\mathcal{H}}_R$ is independent of the network topology and matches the gradient complexity of the centralized near-optimal variance-reduced methods [48–50] for this problem class up to constant factors, under a slightly stronger smoothness assumption; see Remark 3.2.1. Moreover, $\widetilde{\mathcal{H}}_R$ demonstrates a non-asymptotic linear speedup in that the number of component gradient computations required *at each node* to achieve an ϵ-accurate stationary point of F is reduced by a factor of $1/n$, compared to the aforementioned centralized near-optimal algorithms [48–50] that perform all gradient computations at a single node. On the other hand, it is straightforward to verify that $\mathcal{H}_C$ reduces to $\widetilde{\mathcal{H}}_C := \mathcal{O}(N^{1/2}(1-\lambda)^{-1/2}\Delta\epsilon^{-2})$. In other words, in this big-data regime, choosing a large minibatch size $B = \lceil C \rceil$ improves the communication complexity from $\widetilde{\mathcal{K}}_R$ to $\mathcal{K}_C$ while deteriorates the gradient complexity from $\widetilde{\mathcal{H}}_R$ to $\widetilde{\mathcal{H}}_C$, demonstrating an interesting trade-off between computation and communication.

- **Large-scale network regime:** $n = \Omega(N^{1/2}(1-\lambda)^{3/2})$. In this regime, typical to ad hoc IoT networks, i.e., the number of the nodes n and the network spectral gap inverse $(1-\lambda)^{-1}$ are large compared with the total number of samples N, it can be verified that $R = C = 1$ and consequently $\mathcal{H}_R = \mathcal{H}_C$ reduce to $\mathcal{O}(n(1-\lambda)^{-2}\Delta\epsilon^{-2})$ while $\mathcal{K}_R = \mathcal{K}_C$ reduce to $\mathcal{O}((1-\lambda)^{-2}\Delta\epsilon^{-2})$. In other words, in this large-scale network regime, the minibatch size $B = \mathcal{O}(1)$ is preferred since it attains the best possible gradient and communication complexity simultaneously, in the sense of Theorem 3.2.3.

Remark 3.2.5 (Characterization of the big-data regime)**.** We note that the number of nodes n may be interpreted as the intrinsic minibatch size of `GT-SARAH`. We recall that the centralized near-optimal variance-reduced algorithms [48–50] for this problem class retain their best possible gradient complexity if their minibatch size does not exceed $N^{1/2}$ [48]. Thus, the aforementioned big-data regime $n = \mathcal{O}(N^{1/2}(1-\lambda)^3)$

approaches the centralized one as the network connectivity improves and matches the centralized one when the network is fully connected, i.e., $\lambda = 0$.

3.2.5 Numerical experiments

In this section, we illustrate, by numerical experiments, our main theoretical claim that `GT-SARAH` finds a first-order stationary point of Problem (3.1) with a significantly improved gradient complexity compared to the existing decentralized stochastic gradient methods.

3.2.5.1 Setup

We consider a *non-convex* logistic regression model [137] for binary classification over a decentralized network of n nodes with m data samples at each node: $\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\big(f_{i,j}(\mathbf{x}) + r(\mathbf{x})\big)$, such that the logistic loss $f_{i,j}(\mathbf{x})$ and the non-convex regularization $r(\mathbf{x})$ are given by

$$f_{i,j}(\mathbf{x}) := \log\Big[1 + \exp\big\{-(\mathbf{x}^\top \boldsymbol{\theta}_{i,j})\xi_{i,j}\big\}\Big] \qquad \text{and} \qquad r(\mathbf{x}) := R\sum_{d=1}^{p}\frac{[\mathbf{x}]_d^2}{1+[\mathbf{x}]_d^2}, \tag{3.14}$$

where $[\mathbf{x}]_d$ denotes the d-th coordinate of $\mathbf{x}$. In (3.14), note that $\boldsymbol{\theta}_{i,j} \in \mathbb{R}^p$ is the j-th data sample at the i-th node and $\xi_{i,j} \in \{-1, +1\}$ is the corresponding binary label. The details of the datasets under consideration are provided in Table 3.2. We normalize each data sample such that $\|\boldsymbol{\theta}_{i,j}\| = 1, \forall i, j$, and set the regularization parameter as $R = 10^{-3}$. The doubly stochastic weight matrices associated with the networks are generated by the lazy Metroplis rule [27]. We characterize the performance of the algorithms in comparison in terms of the decrease of the network stationary gap versus epochs, where the stationary gap is defined as $\|\nabla F(\overline{\mathbf{x}})\| + \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \overline{\mathbf{x}}\|$, where $\mathbf{x}_i$ is the estimate of the stationary point of F at node i and $\overline{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$, and each epoch represents m component gradient computations at each node.

3.2.5.2 Performance comparisons

We compare the performance of `GT-SARAH` with `DSGT` [4] and `D-GET` [135]; we note that `D2` [3] and `DSGD` [2] are not presented here for conciseness, since in general the former achieves a similar performance with `DSGT` and the latter underperforms `DSGT` and `D2` [3, 15, 31]. Towards the parameter selection of each algorithm, we use the following setup: (i) for `GT-SARAH`, we choose its minibatch size as $B = 1$ and its inner-loop length as $q = m$ in light of Corollary 3.2.1; (ii) for `D-GET`, we choose its minibatch size and inner-loop length as $\lfloor m^{1/2} \rfloor$ under which its convergence is established; see Theorem 1 in [135]; and (iii) we manually optimize the step-sizes for `GT-SARAH`, `D-GET`, and `DSGT` across all experiments.

We first compare the performances of `GT-SARAH`, `DSGT`, and `D-GET` in the big-data regime, that is, the number of samples m at each node is relatively large. To this aim, we distribute the covertype, MiniBooNE,

Table 3.2: Datasets used in numerical experiments, available at https://www.openml.org/.

Dataset	**Number of samples** ($N = nm$)	**dimension** (p)
covertype	100,000	54
MiniBooNE	100,000	51
KDD98	82,000	477
w8a	60,000	300
a9a	48,800	124
Fashion-MNIST (T-shirt versus dress)	10,000	784

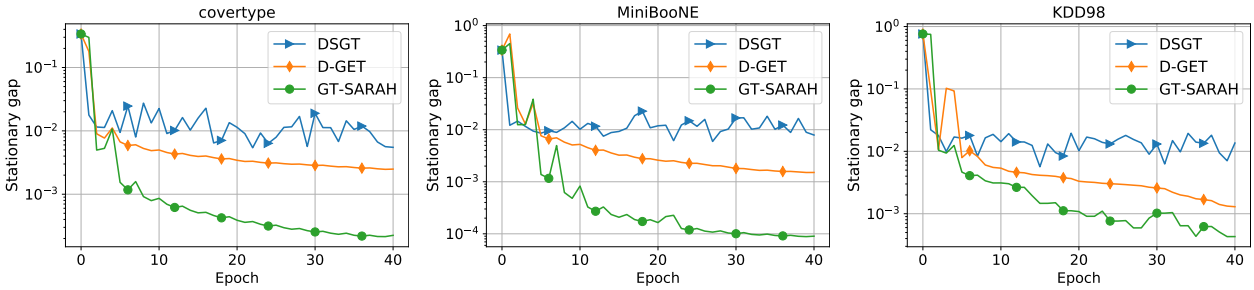

Figure 3.2: Performance comparison of `GT-SARAH`, `DSGT`, and `D-GET` over a 10-node exponential graph on the covertype, MiniBooNE, and KDD98 dataset.

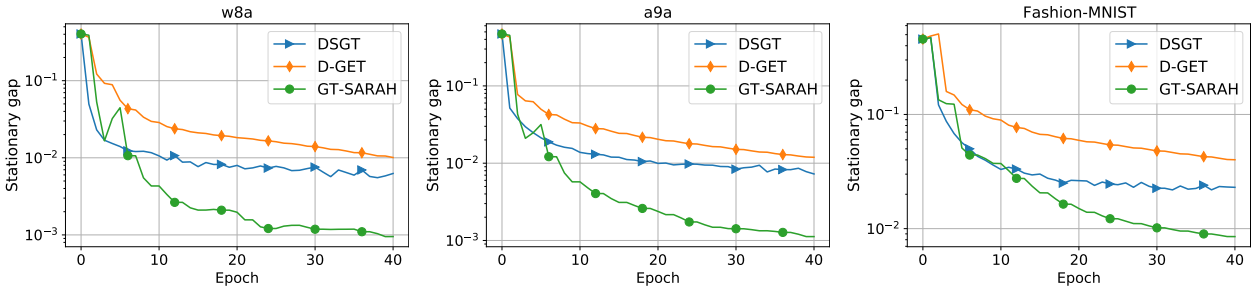

Figure 3.3: Performance comparison of `GT-SARAH`, `DSGT`, and `D-GET` over the 10×10 grid graph on the w8a, a9a, and Fashion-MNIST dataset.

and KDD98 dataset over a 10-node exponential graph [27] whose associated second largest singular value $\lambda \approx 0.71$. The experimental results are presented in Fig. 3.2, where `GT-SARAH` outperforms `DSGT` and `D-GET`. We also observe that `D-GET` outperforms `DSGT` in this case since the performance of the latter is deteriorated by the large variance of the stochastic gradients as the number of the samples m at each node is large.

We next consider the large-scale network regime, where the network spectral gap inverse $(1-\lambda)^{-1}$ and the number of the nodes n are relatively large compared with the local sample size m. We distribute the w8a, a9a, and Fashion-MNIST dataset over the $n = 10 \times 10$ grid graph whose associated second largest eigenvalue $\lambda \approx 0.99$. The performance comparison of the algorithms is shown in Fig. 3.3, where we observe that `GT-SARAH` still outperforms `DSGT` and `D-GET`. Besides, it is worth noting that `D-GET` underperforms `DSGT`

in this case. We provide an explanation about this phenomenon in the following. In the regime where m is relatively small, the variance of the stochastic gradients is relatively small and as a consequence `DSGT` performs well. On the other hand, the minibatch size $\lfloor m^{1/2} \rfloor$ of `D-GET` is too large in this regime to achieve a satisfactory performance; see the related discussion in Subsection 3.2.4.3.

3.2.6 Outline of the convergence analysis

In this section, we present the proof pipeline for Theorems 3.2.1, 3.2.2, and 3.2.3. The analysis framework is novel and general and may be applied to other decentralized algorithms built around variance reduction and gradient tracking. To proceed, we first write `GT-SARAH` in a matrix form. Recall that `GT-SARAH` is a double loop method, where the outer loop index is $s \in \{1, \dots, S\}$ and the inner loop index is $t \in \{0, \dots, q\}$. It is straightforward to verify that `GT-SARAH` can be equivalently written as: $\forall s \geq 1$ and $t \in [0, q]$,

$$\mathbf{y}^{t+1,s} = \mathbf{W}\mathbf{y}^{t,s} + \mathbf{v}^{t,s} - \mathbf{v}^{t-1,s}, \tag{3.15a}$$

$$\mathbf{x}^{t+1,s} = \mathbf{W}\mathbf{x}^{t,s} - \alpha \mathbf{y}^{t+1,s}, \tag{3.15b}$$

where $\mathbf{v}^{t,s}, \mathbf{x}^{t,s}$, and $\mathbf{y}^{t,s}$, in $\mathbb{R}^{np}$, that concatenate local gradient estimators $\{\mathbf{v}_i^{t,s}\}_{i=1}^n$, states $\{\mathbf{x}_i^{t,s}\}_{i=1}^n$, and gradient trackers $\{\mathbf{y}_i^{t,s}\}_{i=1}^n$, respectively, and $\mathbf{W} := \underline{\mathbf{W}} \otimes \mathbf{I}_p$. We recall that $\mathbf{x}^{0,s+1} = \mathbf{x}^{q+1,s}$, $\mathbf{y}^{0,s+1} = \mathbf{y}^{q+1,s}$, $\mathbf{v}^{-1,s+1} = \mathbf{v}^{q,s}, \forall s \geq 1$, and $\mathbf{v}^{-1,1} = \mathbf{0}_{np}$ from Algorithm 3 under the vector notation. Under Assumption 3.2.3, we have [36]

$$\mathbf{J} := \lim_{k \to \infty} \mathbf{W}^k = \left(\tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right) \otimes \mathbf{I}_p,$$

i.e., the power limit of the network weight matrix $\mathbf{W}$ is the exact averaging matrix $\mathbf{J}$. We also introduce the following notation for convenience:

$$\nabla \mathbf{f}(\mathbf{x}^{t,s}) := \left[\nabla f_1(\mathbf{x}_1^{t,s})^\top, \cdots, \nabla f_n(\mathbf{x}_n^{t,s})^\top\right]^\top, \quad \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s}) := \tfrac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\nabla \mathbf{f}(\mathbf{x}^{t,s}),$$
$$\overline{\mathbf{x}}^{t,s} := \tfrac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{x}^{t,s}, \qquad \overline{\mathbf{y}}^{t,s} = \tfrac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{y}^{t,s}, \qquad \overline{\mathbf{v}}^{t,s} := \tfrac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{v}^{t,s}.$$

In particular, we note that $\|\nabla \mathbf{f}(\mathbf{x}^{0,1})\|^2 := \sum_{i=1}^n \|\nabla f_i(\overline{\mathbf{x}}^{0,1})\|^2$. Through the rest of Section 3.2, we assume that Assumptions 3.2.1, 3.2.2, and 3.2.3 hold without explicitly stating them. We define the natural filtration associated with the probability space, an increasing family of sub-σ-algebras of $\mathcal{F}$, as

$$\mathcal{F}^{t,s} := \sigma\left(\sigma\left(\tau_{i,l}^{t-1,s} : i \in \mathcal{V}, l \in [1, B]\right),\ \mathcal{F}^{t-1,s}\right), \quad t \in [2, q+1],\ s \geq 1,$$

where $\mathcal{F}^{1,s} :=: \mathcal{F}^{0,s} := \mathcal{F}^{q+1,s-1}, s \geq 2$, and $\mathcal{F}^{1,1} :=: \mathcal{F}^{0,1}$ are the trivial σ-algebra. It can be verified by induction that $\mathbf{x}^{t,s}$, $\mathbf{y}^{t,s}$ are $\mathcal{F}^{t,s}$-measurable, and $\mathbf{v}^{t,s}$ is $\mathcal{F}^{t+1,s}$-measurable, $\forall s \geq 1$ and $t \in [0, q]$. We assume that the starting point $\overline{\mathbf{x}}^{0,1}$ of `GT-SARAH` is a constant vector. We next present some standard results in

the context of decentralized optimization and gradient tracking methods. The following lemma provides an upper bound on the difference between the exact global gradient and the average of local batch gradients in terms of the state consensus error, as a result of the L-smoothness of each f_i.

Lemma 3.2.1. $\left\|\overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s}) - \nabla F(\overline{\mathbf{x}}^{t,s})\right\|^2 \leq \frac{L^2}{n}\left\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\right\|^2$, $\forall s \geq 1$ *and* $t \in [0, q]$.

Proof. Observe that: $\forall s \geq 1$ and $t \in [0, q]$,

$$\begin{aligned}\left\|\overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s}) - \nabla F(\overline{\mathbf{x}}^{t,s})\right\|^2 &= \frac{1}{n^2}\left\|\sum_{i=1}^n \left(\nabla f_i(\mathbf{x}_i^{t,s}) - \nabla f_i(\overline{\mathbf{x}}^{t,s})\right)\right\|^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}_i^{t,s}) - \nabla f_i(\overline{\mathbf{x}}^{t,s})\right\|^2 \\ &\leq \frac{L^2}{n}\sum_{i=1}^n \left\|\mathbf{x}_i^{t,s} - \overline{\mathbf{x}}^{t,s}\right\|^2,\end{aligned}$$

where the last line is due to the L-smoothness of each f_i. The proof is complete. □

The following are some standard inequalities on the state consensus error.

Lemma 3.2.2. *The following inequalities holds:* $\forall s \geq 1$ *and* $t \in [0, q]$,

$$\left\|\mathbf{x}^{t+1,s} - \mathbf{J}\mathbf{x}^{t+1,s}\right\|^2 \leq \frac{1+\lambda^2}{2}\left\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\right\|^2 + \frac{2\alpha^2}{1-\lambda^2}\left\|\mathbf{y}^{t+1,s} - \mathbf{J}\mathbf{y}^{t+1,s}\right\|^2. \tag{3.16}$$

$$\left\|\mathbf{x}^{t+1,s} - \mathbf{J}\mathbf{x}^{t+1,s}\right\|^2 \leq 2\left\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\right\|^2 + 2\alpha^2\left\|\mathbf{y}^{t+1,s} - \mathbf{J}\mathbf{y}^{t+1,s}\right\|^2. \tag{3.17}$$

Proof. Using (3.15b) and the fact that $\mathbf{JW} = \mathbf{J}$, we have: $\forall s \geq 1$ and $\forall t \in [0, q]$,

$$\begin{aligned}\left\|\mathbf{x}^{t+1,s} - \mathbf{J}\mathbf{x}^{t+1,s}\right\|^2 &= \left\|\mathbf{W}\mathbf{x}^{t,s} - \alpha\mathbf{y}^{t+1,s} - \mathbf{J}\left(\mathbf{W}\mathbf{x}^{t,s} - \alpha\mathbf{y}^{t+1,s}\right)\right\|^2 \\ &= \left\|\mathbf{W}\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s} - \alpha\left(\mathbf{y}^{t+1,s} - \mathbf{J}\mathbf{y}^{t+1,s}\right)\right\|^2 \end{aligned} \tag{3.18}$$

We apply Young's inequality, $\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\eta)\|\mathbf{a}\|^2 + (1+\eta^{-1})\|\mathbf{b}\|^2, \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^{np}$, $\forall \eta > 0$, and Lemma 3.2.4 to (3.18) to obtain: $\forall s \geq 1$ and $\forall t \in [0, q]$,

$$\left\|\mathbf{x}^{t+1,s} - \mathbf{J}\mathbf{x}^{t+1,s}\right\|^2 \leq (1+\eta)\,\lambda^2\left\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\right\|^2 + \left(1+\eta^{-1}\right)\alpha^2\left\|\mathbf{y}^{t+1,s} - \mathbf{J}\mathbf{y}^{t+1,s}\right\|^2.$$

Setting η as $\frac{1-\lambda^2}{2\lambda^2}$ and 1 respectively yields (3.16) and (3.17). □

3.2.6.1 Auxiliary relationships

First, as a consequence of the gradient tracking update (3.15b), it is straightforward to show by induction the following result.

Lemma 3.2.3. $\overline{\mathbf{y}}^{t+1,s} = \overline{\mathbf{v}}^{t,s}, \forall s \geq 1$ *and* $t \in [0, q]$.

Proof. See Section 3.2.7.1. □

The above lemma states that the average of gradient trackers preserves the average of local gradient estimators. Under Assumption 3.2.3, we obtain that the weight matrix $\mathbf{W}$ is a contraction operator [56].

Lemma 3.2.4. $\|\mathbf{W}\mathbf{x} - \mathbf{J}\mathbf{x}\| \leq \lambda \|\mathbf{x} - \mathbf{J}\mathbf{x}\|$, $\forall \mathbf{x} \in \mathbb{R}^{np}$, *for* $\lambda \in [0, 1)$ *defined in* (3.7).

Lemmas 3.2.3 and 3.2.4 are standard in decentralized optimization and gradient tracking [54,56]. The L-smoothness of F leads to the following quadratic upper bound [6]:

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p. \tag{3.19}$$

Consequently, the following descent type lemma on the iterates generated by `GT-SARAH` may be established by setting $\mathbf{y} = \overline{\mathbf{x}}^{t+1,s}$ and $\mathbf{x} = \overline{\mathbf{x}}^{t,s}$ in (3.19) and taking a telescoping sum across all iterations of `GT-SARAH` with the help of Lemmas 3.2.3 and the L-smoothness of each f_i.

Lemma 3.2.5. *If the step-size follows that* $0 < \alpha \leq \frac{1}{2L}$, *then we have:*

$$\begin{aligned}\mathbb{E}\big[F\big(\overline{\mathbf{x}}^{q+1,S}\big)\big] \leq F\big(\overline{\mathbf{x}}^{0,1}\big) - \frac{\alpha}{2}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\nabla F(\overline{\mathbf{x}}^{t,s})\big\|^2\Big] - \frac{\alpha}{4}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\big[\|\overline{\mathbf{v}}^{t,s}\|^2\big] \\ + \alpha\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s})\big\|^2\Big] + \alpha L^2 \sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\frac{\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\|^2}{n}\right].\end{aligned}$$

Proof. See Section 3.2.7.2. □

In light of Lemma 3.2.5, our analysis approach is to derive the range of α of `GT-SARAH` such that

$$\frac{1}{4}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\big[\|\overline{\mathbf{v}}^{t,s}\|^2\big] - \sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s})\big\|^2\Big] - L^2\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\frac{\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\|^2}{n}\right] \geq 0,$$

and therefore establishes the convergence of `GT-SARAH` to a first-order stationary point following the standard arguments in *batch* gradient descent for non-convex problems [6, 7]. To this aim, we need to derive upper bounds for two error terms in the above expression: (i) $\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s})\|^2$, the gradient estimation error; and (ii) $\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\|^2$, the state consensus error. We quantify these two errors next and then return to Lemma 3.2.5. The following lemma is obtained with similar probabilistic arguments for `SARAH`-type [48–50] estimators, however, with subtle modifications due to the decentralized network effect.

Lemma 3.2.6. *We have:* $\forall s \geq 1$,

$$\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s})\big\|^2\Big] \leq \frac{3q\alpha^2 L^2}{nB}\sum_{t=0}^{q-1}\mathbb{E}\big[\|\overline{\mathbf{v}}^{t,s}\|^2\big] + \frac{6qL^2}{nB}\sum_{t=0}^{q}\mathbb{E}\left[\frac{\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\|^2}{n}\right].$$

Proof. See Section 3.2.7.3. □

Note that Lemma 3.2.6 shows that the accumulated gradient estimation error over one inner loop may be bounded by the accumulated state consensus error and the norm of the gradient estimators. Lemma 3.2.6 thus may be used to simplify the right hand side of the descent inequality in Lemma 3.2.5. Naturally, what is left is to seek an upper bound for the state consensus error in terms of $\mathbb{E}[\|\overline{\mathbf{v}}^{t,s}\|^2]$. This result is presented in the following lemma.

Lemma 3.2.7. *If the step-size follows* $0 < \alpha \leq \frac{(1-\lambda^2)^2}{8\sqrt{42}L}$, *then*

$$\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\frac{\|\mathbf{x}^{t,s}-\mathbf{J}\mathbf{x}^{t,s}\|^2}{n}\right] \leq \frac{64\alpha^2}{(1-\lambda^2)^3}\frac{\|\nabla\mathbf{f}(\mathbf{x}^{0,1})\|^2}{n} + \frac{1536\alpha^4L^2}{(1-\lambda^2)^4}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\big[\|\overline{\mathbf{v}}^{t,s}\|^2\big].$$

Proof. See Section 3.2.7.4. □

Establishing Lemma 3.2.7 requires a careful analysis; here, we provide a brief sketch. Recall the `GT-SARAH` algorithm in (3.15a)-(3.15b) and note that the state vector $\mathbf{x}^{t,s}$ is coupled with the gradient tracker $\mathbf{y}^{t,s}$. Thus, in order to quantify the state consensus error $\|\mathbf{x}^{t,s}-\mathbf{J}\mathbf{x}^{t,s}\|^2$, we need to establish its relationship with the gradient tracking error $\|\mathbf{y}^{t,s}-\mathbf{J}\mathbf{y}^{t,s}\|^2$. In fact, we show that these coupled errors jointly formulate a linear time-invariant (LTI) system dynamics whose system matrix is stable under a certain range of the step-size α. Solving this LTI yields Lemma 3.2.7.

Finally, it is straightforward to use Lemmas 3.2.6 and 3.2.7 to refine the descent inequality in Lemma 3.2.5 to obtain the following result.

Lemma 3.2.8. *If* $0 < \alpha \leq \overline{\alpha} := \min\left\{\frac{(1-\lambda^2)^2}{4\sqrt{42}}, \left(\frac{nB}{6q}\right)^{1/2}, \left(\frac{4nB}{7nB+24q}\right)^{1/4}\frac{1-\lambda^2}{6}\right\}\frac{1}{2L}$, *then*

$$L^2\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\frac{\|\mathbf{x}^{t,s}-\mathbf{J}\mathbf{x}^{t,s}\|^2}{n}\right] + \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\nabla F\big(\mathbf{x}_i^{t,s}\big)\big\|^2\Big]$$
$$\leq \frac{4\big(F\big(\overline{\mathbf{x}}^{0,1}\big)-F^*\big)}{\alpha} + \left(\frac{7}{4}+\frac{6q}{nB}\right)\frac{256\alpha^2L^2}{(1-\lambda^2)^3}\frac{\|\nabla\mathbf{f}(\mathbf{x}^{0,1})\|^2}{n}.$$

Proof. See Section 3.2.7.5. □

We note that the descent inequality in Lemma 3.2.8 that characterizes the convergence of **`GT-SARAH`** is independent of the variance of local gradient estimators and of the difference between the local and the global gradient. In fact, it has similarities to that of the centralized *batch* gradient descent [6, 7]; see also the discussion on `DSGD` in Section 3.1. This is a consequence of the joint use of the local variance reduction and the global gradient tracking. This is essentially why we are able to match the gradient complexity of the centralized near-optimal methods for finite sum problems and obtain the almost sure convergence guarantee of `GT-SARAH` to a stationary point.

3.2.6.2 Proofs of the main theorems

With Lemma 3.2.8 at hand, Theorems 3.2.1, 3.2.2, and 3.2.3 are now straightforward to prove.

Proof of Theorem 3.2.1. We observe from Lemma 3.2.8 that if $0 < \alpha \leq \overline{\alpha}$, then

$$\sum_{s=1}^{\infty}\sum_{t=0}^{q}\mathbb{E}\big[\|\nabla F(\mathbf{x}_i^{t,s})\|^2 + L^2\|\mathbf{x}_i^{t,s} - \overline{\mathbf{x}}^{t,s}\|^2\big] < \infty, \qquad \forall i \in \mathcal{V},$$

which implies all nodes achieve consensus and converge to a stationary point in the mean-squared sense. Further, by monotone convergence theorem [127], we exchange the order of the expectation and the series to obtain:

$$\mathbb{E}\left[\sum_{s=1}^{\infty}\sum_{t=0}^{q}\left(\|\nabla F(\mathbf{x}_i^{t,s})\|^2 + L^2\|\mathbf{x}_i^{t,s} - \overline{\mathbf{x}}^{t,s}\|^2\right)\right] < \infty, \qquad \forall i \in \mathcal{V},$$

which leads to

$$\mathbb{P}\left(\sum_{s=1}^{\infty}\sum_{t=0}^{q}\left(\|\nabla F(\mathbf{x}_i^{t,s})\|^2 + L^2\|\mathbf{x}_i^{t,s} - \overline{\mathbf{x}}^{t,s}\|^2\right) < \infty\right) = 1, \qquad \forall i \in \mathcal{V},$$

i.e., the consensus and convergence to a stationary point in the almost sure sense. □

Proof of Theorem 3.2.2. We recall the metric of the outer loop complexity in Definition 3.2.1 and we divide the descent inequality in Lemma 3.2.8 by $S(q+1)$ from both sides. It is then clear that to find an ϵ-accurate stationary point of F, it suffices to choose the total number of the outer loop iterations S such that

$$\frac{4\left(F(\overline{\mathbf{x}}^{0,1}) - F^*\right)}{S(q+1)\alpha} + \left(\frac{7}{4} + \frac{6q}{nB}\right)\frac{256\alpha^2L^2}{S(q+1)(1-\lambda^2)^3}\frac{\|\nabla\mathbf{f}(\mathbf{x}^{0,1})\|^2}{n} \leq \epsilon^2. \tag{3.20}$$

The proof follows by that if $0 < \alpha \leq \left(\frac{4nB}{7nB+24q}\right)^{1/3}\frac{1-\lambda^2}{12L}$, then $(\frac{7}{4} + \frac{6q}{nB})\frac{256\alpha^2L^2}{(1-\lambda^2)^3} \leq \frac{1}{\alpha L}$, and by solving for the lower bound on S such that (3.20) holds. □

Proof of Theorem 3.2.3. During each inner loop, `GT-SARAH` incurs $n(m+2qB)$ component gradient computations across all nodes and q rounds of communication of the network. Hence, to find an ϵ-accurate stationary point of F, `GT-SARAH` requires, according to Theorem 3.2.2, at most

$$\mathcal{H} = \mathcal{O}\left(\frac{n(m+qB)}{q\alpha L\epsilon^2}\left(L\left(F(\overline{\mathbf{x}}^{0,1}) - F^*\right) + \frac{\|\nabla\mathbf{f}(\mathbf{x}^{0,1})\|^2}{n}\right)\right)$$

component gradient computations across all nodes and

$$\mathcal{K} = \mathcal{O}\left(\frac{1}{\alpha L\epsilon^2}\left(L\left(F(\overline{\mathbf{x}}^{0,1}) - F^*\right) + \frac{\|\nabla\mathbf{f}(\mathbf{x}^{0,1})\|^2}{n}\right)\right)$$

rounds of communication of the network. The proof follows by setting the step-size α as its upper bound in Theorem 3.2.2 and the length of the inner loop as $q = \mathcal{O}(\frac{m}{B})$. □

3.2.7 Detailed proofs for lemmata in Section 3.2.6

In this section, we present the proofs of the technical lemmas 3.2.3, 3.2.5, 3.2.6, 3.2.7, 3.2.8.

3.2.7.1 Proof of Lemma 3.2.3

Using Assumption 3.2.3, we multiply (3.15a) by $\frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)$ to obtain: $\forall s \geq 1$ and $t \in [0, q]$,

$$\begin{aligned}
\overline{\mathbf{y}}^{t+1,s} &= \overline{\mathbf{y}}^{t,s} + \overline{\mathbf{v}}^{t,s} - \overline{\mathbf{v}}^{t-1,s} \\
&= \overline{\mathbf{y}}^{t-1,s} + \overline{\mathbf{v}}^{t,s} - \overline{\mathbf{v}}^{t-2,s} \\
&\cdots \\
&= \overline{\mathbf{y}}^{0,s} + \overline{\mathbf{v}}^{t,s} - \overline{\mathbf{v}}^{-1,s} \\
&= \overline{\mathbf{y}}^{q+1,s-1} + \overline{\mathbf{v}}^{t,s} - \overline{\mathbf{v}}^{q,s-1} \\
&\cdots \\
&= \overline{\mathbf{y}}^{0,1} + \overline{\mathbf{v}}^{t,s} - \overline{\mathbf{v}}^{-1,1} = \overline{\mathbf{v}}^{t,s},
\end{aligned}$$

where the above series of equalities follows directly from the updates of `GT-SARAH`.

3.2.7.2 Proof of Lemma 3.2.5

We multiply (3.15b) by $\frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)$ and then use Lemma 3.2.3 to obtain the recursion of the mean state $\overline{\mathbf{x}}^{t,s}$ as follows:

$$\overline{\mathbf{x}}^{t+1,s} = \overline{\mathbf{x}}^{t,s} - \alpha \overline{\mathbf{y}}^{t+1,s} = \overline{\mathbf{x}}^{t,s} - \alpha \overline{\mathbf{v}}^{t,s}, \qquad \forall s \geq 1 \text{ and } t \in [0, q].$$

Setting $\mathbf{y} = \overline{\mathbf{x}}^{t+1,s}$ and $\mathbf{x} = \overline{\mathbf{x}}^{t,s}$ in (3.19), we have: $\forall s \geq 1$ and $t \in [0, q]$,

$$F(\overline{\mathbf{x}}^{t+1,s}) \leq F(\overline{\mathbf{x}}^{t,s}) - \alpha \left\langle \nabla F(\overline{\mathbf{x}}^{t,s}), \overline{\mathbf{v}}^{t,s} \right\rangle + \tfrac{\alpha^2 L}{2} \left\| \overline{\mathbf{v}}^{t,s} \right\|^2. \tag{3.21}$$

Applying $\langle \mathbf{a}, \mathbf{b} \rangle = 0.5 \left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2 \right), \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, to (3.21), we obtain an inequality that characterizes the descent of the network mean state over one inner loop iteration: $\forall s \geq 1$ and $t \in [0, q]$,

$$\begin{aligned}
F(\overline{\mathbf{x}}^{t+1,s}) &\leq F(\overline{\mathbf{x}}^{t,s}) - \tfrac{\alpha}{2} \left\| \nabla F(\overline{\mathbf{x}}^{t,s}) \right\|^2 - \tfrac{\alpha(1-\alpha L)}{2} \left\| \overline{\mathbf{v}}^{t,s} \right\|^2 + \tfrac{\alpha}{2} \left\| \overline{\mathbf{v}}^{t,s} - \nabla F(\overline{\mathbf{x}}^{t,s}) \right\|^2, \\
&\leq F(\overline{\mathbf{x}}^{t,s}) - \tfrac{\alpha}{2} \left\| \nabla F(\overline{\mathbf{x}}^{t,s}) \right\|^2 - \tfrac{\alpha}{4} \left\| \overline{\mathbf{v}}^{t,s} \right\|^2 + \alpha \left\| \overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s}) \right\|^2 \\
&\quad + \alpha \left\| \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s}) - \nabla F(\overline{\mathbf{x}}^{t,s}) \right\|^2 \qquad (3.22) \\
&\leq F(\overline{\mathbf{x}}^{t,s}) - \tfrac{\alpha}{2} \left\| \nabla F(\overline{\mathbf{x}}^{t,s}) \right\|^2 - \tfrac{\alpha}{4} \left\| \overline{\mathbf{v}}^{t,s} \right\|^2 + \alpha \left\| \overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s}) \right\|^2 \\
&\quad + \tfrac{\alpha L^2}{n} \left\| \mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s} \right\|^2, \qquad (3.23)
\end{aligned}$$

where (3.22) is due to $0 < \alpha \leq \frac{1}{2L}$ and (3.23) is due to Lemma 3.2.1. We then take the telescoping sum of (3.23) over t from 0 to q to obtain: $\forall s \geq 1$,

$$\begin{aligned}
F(\overline{\mathbf{x}}^{0,s+1}) \leq F(\overline{\mathbf{x}}^{0,s}) &- \tfrac{\alpha}{2} \textstyle\sum_{t=0}^{q} \|\nabla F(\overline{\mathbf{x}}^{t,s})\|^2 - \tfrac{\alpha}{4} \textstyle\sum_{t=0}^{q} \left\| \overline{\mathbf{v}}^{t,s} \right\|^2 \qquad (3.24) \\
&+ \alpha \textstyle\sum_{t=0}^{q} \left\| \overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s}) \right\|^2 + \tfrac{\alpha L^2}{n} \textstyle\sum_{t=0}^{q} \left\| \mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s} \right\|^2.
\end{aligned}$$

The proof then follows by taking the telescoping sum of (3.24) over s from 1 to S and taking the expectation of the resulting inequality.

3.2.7.3 Proof of Lemma 3.2.6

We first provide a useful result.

Lemma 3.2.9. *The following inequality holds:* $\forall s \geq 1$, $\forall t \in [1, q]$, $\forall i \in \mathcal{V}$, $\forall l \in [1, B]$,

$$\mathbb{E}\Big[\Big\|\nabla f_{i,\tau_{i,l}^{t,s}}\big(\mathbf{x}_i^{t,s}\big) - \nabla f_{i,\tau_{i,l}^{t,s}}\big(\mathbf{x}_i^{t-1,s}\big)\Big\|^2 \big|\mathcal{F}^{t,s}\Big] \leq L^2\big\|\mathbf{x}_i^{t,s} - \mathbf{x}_i^{t-1,s}\big\|^2.$$

Proof. In the following, we denote $\mathbb{1}\{A\}$ as the indicator function of an event $A \in \mathcal{F}$. Observe that: $\forall s \geq 1$, $\forall t \in [1, q]$, $\forall i \in \mathcal{V}$, $\forall l \in [1, B]$,

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\nabla f_{i,\tau_{i,l}^{t,s}}\big(\mathbf{x}_i^{t,s}\big) - \nabla f_{i,\tau_{i,l}^{t,s}}\big(\mathbf{x}_i^{t-1,s}\big)\Big\|^2\Big|\mathcal{F}^{t,s}\Big] \\
&= \mathbb{E}\Bigg[\Bigg\|\sum_{j=1}^{m}\mathbb{1}\big\{\tau_{i,l}^{t,s} = j\big\}\Big(\nabla f_{i,j}\big(\mathbf{x}_i^{t,s}\big) - \nabla f_{i,j}\big(\mathbf{x}_i^{t-1,s}\big)\Big)\Bigg\|^2\Bigg|\mathcal{F}^{t,s}\Bigg] \\
&= \sum_{j=1}^{m}\mathbb{E}\Big[\mathbb{1}\big\{\tau_{i,l}^{t,s} = j\big\}\Big\|\nabla f_{i,j}\big(\mathbf{x}_i^{t,s}\big) - \nabla f_{i,j}\big(\mathbf{x}_i^{t-1,s}\big)\Big\|^2\Big|\mathcal{F}^{t,s}\Big] \\
&= \sum_{j=1}^{m}\mathbb{E}\Big[\mathbb{1}\big\{\tau_{i,l}^{t,s} = j\big\}|\mathcal{F}\Big]\Big\|\nabla f_{i,j}\big(\mathbf{x}_i^{t,s}\big) - \nabla f_{i,j}\big(\mathbf{x}_i^{t-1,s}\big)\Big\|^2 \\
&= \frac{1}{m}\sum_{j=1}^{m}\Big\|\nabla f_{i,j}\big(\mathbf{x}_i^{t,s}\big) - \nabla f_{i,j}\big(\mathbf{x}_i^{t-1,s}\big)\Big\|^2,
\end{aligned}$$

where the last line uses that $\tau_{i,l}^{t,s}$ is independent of $\mathcal{F}$, i.e., $\mathbb{E}\big[\mathbb{1}\{\tau_{i,l}^{t,s} = j\}|\mathcal{F}\big] = \frac{1}{m}$. The proof follows by using Assumption 3.2.1. □

Next, we derive an upper bound on the estimation error of the average of local `SARAH` gradient estimators across the nodes at each inner loop iteration.

Lemma 3.2.10. *The following inequality holds:* $\forall s \geq 1$ *and* $t \in [1, q]$,

$$\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla\mathbf{f}}(\mathbf{x}^{t,s})\big\|^2\Big] \leq \frac{3\alpha^2L^2}{nB}\sum_{u=0}^{t-1}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{u,s}\big\|^2\Big] + \frac{6L^2}{n^2B}\sum_{u=0}^{t}\mathbb{E}\Big[\big\|\mathbf{x}^{u,s} - \mathbf{J}\mathbf{x}^{u,s}\big\|^2\Big].$$

Proof. For the ease of exposition, we denote: $\forall t \in [1, q]$, $\forall s \geq 1$, $\forall i \in \mathcal{V}$, $\forall l \in [1, B]$,

$$\widehat{\nabla}_{i,l}^{t,s} := \nabla f_{i,\tau_{i,l}^{t,s}}\big(\mathbf{x}_i^{t,s}\big) - \nabla f_{i,\tau_{i,l}^{t,s}}\big(\mathbf{x}_i^{t-1,s}\big), \qquad \widehat{\nabla}_i^{t,s} := \tfrac{1}{B}\textstyle\sum_{l=1}^{B}\widehat{\nabla}_{i,l}^{t,s}. \tag{3.25}$$

Since $\mathbf{x}_i^{t,s}$ and $\mathbf{x}_i^{t-1,s}$ are $\mathcal{F}$-measurable, we have

$$\mathbb{E}\Big[\widehat{\nabla}_{i,l}^{t,s}|\mathcal{F}\Big] = \mathbb{E}\Big[\widehat{\nabla}_i^{t,s}|\mathcal{F}\Big] = \nabla f_i\big(\mathbf{x}_i^{t,s}\big) - \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big). \tag{3.26}$$

With the notations in (3.25), the local update of $\mathbf{v}_i^{t,s}$ described in Algorithm 3 may be written as

$$\mathbf{v}_i^{t,s} = \widehat{\nabla}_i^{t,s} + \mathbf{v}_i^{t-1,s}, \qquad \forall t \in [1,q],\ \forall s \geq 1,\ \forall i \in \mathcal{V}.$$

In the light of (3.26), we have the following: $\forall s \geq 1$ and $t \in [1,q]$,

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s})\big\|^2 \big| \mathcal{F}^{t,s}\Big] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \Big(\widehat{\nabla}_i^{t,s} + \mathbf{v}_i^{t-1,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big)\Big)\right\|^2 \Big| \mathcal{F}^{t,s}\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \Big(\widehat{\nabla}_i^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big) + \mathbf{v}_i^{t-1,s} - \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\Big)\right\|^2 \Big| \mathcal{F}^{t,s}\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \Big(\widehat{\nabla}_i^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\Big)\right\|^2 \Big| \mathcal{F}^{t,s}\right] + \left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{v}_i^{t-1,s} - \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\Big)\right\|^2 \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \Big(\widehat{\nabla}_i^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\Big)\right\|^2 \Big| \mathcal{F}^{t,s}\right] + \big\|\overline{\mathbf{v}}^{t-1,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t-1,s})\big\|^2, \qquad (3.27)
\end{aligned}$$

where the third equality is due to (3.26) and the fact that $\sum_{i=1}^n \big(\mathbf{v}_i^{t-1,s} - \nabla f_i(\mathbf{x}_i^{t-1,s})\big)$ is $\mathcal{F}^{t,s}$-measurable. To proceed from (3.27), we note that since the collection of random variables $\big\{\tau_{i,l}^{t,s} : i \in \mathcal{V}, l \in [1,B]\big\}$ are independent of each other and of the filtration $\mathcal{F}^{t,s}$, by (3.26), we have: $\forall t \in [1,q]$ and $s \geq 1$,

$$\mathbb{E}\Big[\Big\langle \widehat{\nabla}_i^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big), \widehat{\nabla}_r^{t,s} - \nabla f_r\big(\mathbf{x}_r^{t,s}\big) + \nabla f_r\big(\mathbf{x}_r^{t-1,s}\big)\Big\rangle \Big| \mathcal{F}^{t,s}\Big] = 0, \qquad (3.28)$$

whenever $i, r \in \mathcal{V}$ such that $i \neq r$. Similarly, we have: $\forall t \in [1,q]$ and $s \geq 1$, $\forall i \in \mathcal{V}$,

$$\mathbb{E}\Big[\Big\langle \widehat{\nabla}_{i,l}^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big), \widehat{\nabla}_{i,h}^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\Big\rangle \Big| \mathcal{F}^{t,s}\Big] = 0, \qquad (3.29)$$

whenever $l, h \in [1,m]$ such that $l \neq h$. With the help of (3.28) and (3.29), we may simplify (3.27) in the following: $\forall s \geq 1$ and $t \in [1,q]$,

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s})\big\|^2 \big| \mathcal{F}^{t,s}\Big] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\widehat{\nabla}_i^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\right\|^2 \Big| \mathcal{F}^{t,s}\right] + \big\|\overline{\mathbf{v}}^{t-1,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t-1,s})\big\|^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\frac{1}{B}\sum_{l=1}^B \Big(\widehat{\nabla}_{i,l}^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\Big)\right\|^2 \Big| \mathcal{F}^{t,s}\right] + \big\|\overline{\mathbf{v}}^{t-1,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t-1,s})\big\|^2 \\
&= \frac{1}{(nB)^2}\sum_{i=1}^n\sum_{l=1}^B \mathbb{E}\left[\left\|\widehat{\nabla}_{i,l}^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\right\|^2 \Big| \mathcal{F}^{t,s}\right] + \big\|\overline{\mathbf{v}}^{t-1,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t-1,s})\big\|^2, \qquad (3.30)
\end{aligned}$$

where the first line is due to (3.28) and the last line is due to (3.29). To proceed from (3.30), we observe

that $\forall t \in [1,q]$, $\forall s \geq 1$, $\forall i \in \mathcal{V}$, $\forall l \in [1,B]$,

$$\begin{aligned}\mathbb{E}\Big[\Big\|\widehat{\nabla}_{i,l}^{t,s} - \nabla f_i(\mathbf{x}_i^{t,s}) + \nabla f_i(\mathbf{x}_i^{t-1,s})\Big\|^2\Big|\mathcal{F}^{t,s}\Big] &= \mathbb{E}\Big[\Big\|\widehat{\nabla}_{i,l}^{t,s} - \mathbb{E}\Big[\widehat{\nabla}_{i,l}^{t,s}|\mathcal{F}\Big]\Big\|^2|\mathcal{F}^{t,s}\Big] \\ &\leq \mathbb{E}\Big[\Big\|\widehat{\nabla}_{i,l}^{t,s}\Big\|^2|\mathcal{F}^{t,s}\Big] \\ &\leq L^2\big\|\mathbf{x}_i^{t,s} - \mathbf{x}_i^{t-1,s}\big\|^2, \end{aligned} \tag{3.31}$$

where the last line uses Lemma 3.2.9. Applying (3.31) to (3.30) yields: $\forall s \geq 1, t \in [1,q]$,

$$\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s})\big\|^2|\mathcal{F}^{t,s}\Big] \leq \frac{L^2}{n^2B}\big\|\mathbf{x}^{t,s} - \mathbf{x}^{t-1,s}\big\|^2 + \big\|\overline{\mathbf{v}}^{t-1,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t-1,s})\big\|^2. \tag{3.32}$$

We next bound the first term on the right hand side of (3.32). Observe that $\forall s \geq 1$ and $t \in [1, q+1]$,

$$\begin{aligned}\big\|\mathbf{x}^{t,s} - \mathbf{x}^{t-1,s}\big\|^2 &= \big\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s} + \mathbf{J}\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t-1,s} + \mathbf{J}\mathbf{x}^{t-1,s} - \mathbf{x}^{t-1,s}\big\|^2 \\ &\leq 3\big\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\big\|^2 + 3n\big\|\overline{\mathbf{x}}^{t,s} - \overline{\mathbf{x}}^{t-1,s}\big\|^2 + 3\big\|\mathbf{x}^{t-1,s} - \mathbf{J}\mathbf{x}^{t-1,s}\big\|^2 \\ &= 3\big\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\big\|^2 + 3n\alpha^2\big\|\overline{\mathbf{v}}^{t-1,s}\big\|^2 + 3\big\|\mathbf{x}^{t-1,s} - \mathbf{J}\mathbf{x}^{t-1,s}\big\|^2. \end{aligned} \tag{3.33}$$

Applying (3.33) to (3.32) and taking the expectation of the resulting inequality leads to: $\forall s \geq 1$ and $t \in [1,q]$,

$$\begin{aligned}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s})\big\|^2\Big] \leq{}& \mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t-1,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t-1,s})\big\|^2\Big] + \frac{3\alpha^2L^2}{nB}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t-1,s}\big\|^2\Big] \\ &+ \frac{3L^2}{n^2B}\mathbb{E}\Big[\big\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\big\|^2\Big] + \frac{3L^2}{n^2B}\mathbb{E}\Big[\big\|\mathbf{x}^{t-1,s} - \mathbf{J}\mathbf{x}^{t-1,s}\big\|^2\Big]. \end{aligned} \tag{3.34}$$

We recall the initialization of each inner loop that $\overline{\mathbf{v}}^{0,s} = \overline{\nabla \mathbf{f}}(\mathbf{x}^{0,s}), \forall s \geq 1$, and take the telescoping sum of (3.34) over t from 1 to z to obtain: $\forall s \geq 1$ and $\forall z \in [1,q]$,

$$\begin{aligned}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{z,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{z,s})\big\|^2\Big] \leq{}& \frac{3\alpha^2L^2}{nB}\sum_{t=1}^{z}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t-1,s}\big\|^2\Big] + \frac{3L^2}{n^2B}\sum_{t=1}^{z}\mathbb{E}\Big[\big\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\big\|^2\Big] \\ &+ \frac{3L^2}{n^2B}\sum_{t=1}^{z}\mathbb{E}\Big[\big\|\mathbf{x}^{t-1,s} - \mathbf{J}\mathbf{x}^{t-1,s}\big\|^2\Big]. \end{aligned} \tag{3.35}$$

The proof follows by merging the last two terms on the right hand side of (3.35). □

Proof of Lemma 3.2.6. Summing up Lemma 3.2.10 over t from 1 to q gives: $\forall s \geq 1$,

$$\sum_{t=1}^{q}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla \mathbf{f}}(\mathbf{x}^{t,s})\big\|^2\Big] \leq \frac{3\alpha^2L^2}{nB}\sum_{t=1}^{q}\sum_{u=0}^{t-1}\mathbb{E}\big[\|\overline{\mathbf{v}}^{u,s}\|^2\big] + \frac{6L^2}{n^2B}\sum_{t=1}^{q}\sum_{u=0}^{t}\mathbb{E}\Big[\big\|\mathbf{x}^{u,s} - \mathbf{J}\mathbf{x}^{u,s}\big\|^2\Big]. \tag{3.36}$$

The proof follows by relaxing the right hand side of (3.36) on the summations and the initialization of each inner loop that $\overline{\mathbf{v}}^{0,s} = \overline{\nabla \mathbf{f}}(\mathbf{x}^{0,s}), \forall s \geq 1$. □

3.2.7.4 Proof of Lemma 3.2.7

We first provide some useful bounds on the gradient estimator tracking errors. These bounds will later be coupled with (3.16) to formulate a dynamical system to characterize the error evolution of `GT-SARAH`. The

following lemma establishes an upper bound on the sum of the local gradient estimation errors across the nodes. Its proof is similar to that of Lemma 3.2.10.

Lemma 3.2.11. *The following inequality holds* $\forall s \geq 1$ *and* $t \in [1, q]$,

$$\mathbb{E}\Big[\big\|\mathbf{v}^{t,s} - \nabla\mathbf{f}(\mathbf{x}^{t,s})\big\|^2\Big] \leq \frac{3n\alpha^2L^2}{B}\sum_{u=0}^{t-1}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{u,s}\big\|^2\Big] + \frac{6L^2}{B}\sum_{u=0}^{t}\mathbb{E}\Big[\big\|\mathbf{x}^{u,s} - \mathbf{J}\mathbf{x}^{u,s}\big\|^2\Big].$$

Proof. See Section 3.2.7.6. □

We note that Lemma 3.2.11 does not follow directly from the results of Lemma 3.2.6 because $\mathbf{v}_i^{t,s}$ is *not* a conditionally unbiased estimator of $\nabla f_i(\mathbf{x}_i^{t,s})$ with respect to $\mathcal{F}^{t,s}$. With Lemma 3.2.11 at hand, we now quantify the gradient tracking errors.

Lemma 3.2.12. *We have the following three statements.*

(i) It holds that $\big\|\mathbf{y}^{1,1} - \mathbf{J}\mathbf{y}^{1,1}\big\|^2 \leq \big\|\nabla\mathbf{f}\big(\mathbf{x}^{0,1}\big)\big\|^2$.

(ii) If $0 < \alpha \leq \frac{1-\lambda^2}{4\sqrt{3}L}$, *the following inequality holds:* $\forall s \geq 1$ *and* $t \in [1, q]$,

$$\mathbb{E}\left[\frac{\|\mathbf{y}^{t+1,s} - \mathbf{J}\mathbf{y}^{t+1,s}\|^2}{nL^2}\right] \leq \frac{3+\lambda^2}{4}\mathbb{E}\left[\frac{\|\mathbf{y}^{t,s} - \mathbf{J}\mathbf{y}^{t,s}\|^2}{nL^2}\right] + \frac{18}{1-\lambda^2}\mathbb{E}\left[\frac{\|\mathbf{x}^{t-1,s} - \mathbf{J}\mathbf{x}^{t-1,s}\|^2}{n}\right] + \frac{6\alpha^2}{1-\lambda^2}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t-1,s}\big\|^2\Big].$$

(iii) If $0 < \alpha \leq \frac{1-\lambda^2}{4\sqrt{6}L}$, *the following inequality holds:* $\forall s \geq 2$,

$$\mathbb{E}\left[\frac{\|\mathbf{y}^{1,s} - \mathbf{J}\mathbf{y}^{1,s}\|^2}{nL^2}\right] \leq \frac{3+\lambda^2}{4}\mathbb{E}\left[\frac{\|\mathbf{y}^{q+1,s-1} - \mathbf{J}\mathbf{y}^{q+1,s-1}\|^2}{nL^2}\right] + \frac{18}{1-\lambda^2}\mathbb{E}\left[\frac{\|\mathbf{x}^{q,s-1} - \mathbf{J}\mathbf{x}^{q,s-1}\|^2}{n}\right] + \frac{12\alpha^2}{1-\lambda^2}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s-1}\big\|^2\Big] + \frac{42}{1-\lambda^2}\sum_{t=0}^{q}\mathbb{E}\left[\frac{\|\mathbf{x}^{t,s-1} - \mathbf{J}\mathbf{x}^{t,s-1}\|^2}{n}\right].$$

Proof. ***(i)*** Recall that $\mathbf{v}^{-1,1} = \mathbf{0}_{np}$, $\mathbf{y}^{0,1} = \mathbf{0}_{np}$ and $\mathbf{v}^{0,1} = \nabla\mathbf{f}\big(\mathbf{x}^{0,1}\big)$. Using the gradient tracking update at iteration $(1,1)$ and $\|\mathbf{I}_{np} - \mathbf{J}\| = 1$, we have:

$$\big\|\mathbf{y}^{1,1} - \mathbf{J}\mathbf{y}^{1,1}\big\|^2 = \big\|(\mathbf{I}_{np} - \mathbf{J})\left(\mathbf{W}\mathbf{y}^{0,1} + \mathbf{v}^{0,1} - \mathbf{v}^{-1,1}\right)\big\|^2 \leq \big\|\nabla\mathbf{f}\big(\mathbf{x}^{0,1}\big)\big\|^2,$$

which proves the first statement in the lemma. In the following, we prove the second and the third statements. We have: $\forall s \geq 1$ and $\forall t \in [0, q]$,

$$\begin{aligned}\big\|\mathbf{y}^{t+1,s} - \mathbf{J}\mathbf{y}^{t+1,s}\big\|^2 &= \big\|\mathbf{W}\mathbf{y}^{t,s} + \mathbf{v}^{t,s} - \mathbf{v}^{t-1,s} - \mathbf{J}\left(\mathbf{W}\mathbf{y}^{t,s} + \mathbf{v}^{t,s} - \mathbf{v}^{t-1,s}\right)\big\|^2 \\ &= \big\|\mathbf{W}\mathbf{y}^{t,s} - \mathbf{J}\mathbf{y}^{t,s} + (\mathbf{I}_{np} - \mathbf{J})\left(\mathbf{v}^{t,s} - \mathbf{v}^{t-1,s}\right)\big\|^2. \end{aligned} \tag{3.37}$$

We apply the inequality that $\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\eta)\|\mathbf{a}\|^2 + (1+\frac{1}{\eta})\|\mathbf{b}\|^2$, $\forall \mathbf{a},\mathbf{b}\in\mathbb{R}^{np}$, with $\eta = \frac{1-\lambda^2}{2\lambda^2}$ and that $\|\mathbf{I}_{np}-\mathbf{J}\|=1$ to (3.37) to obtain: $\forall s\geq 1$ and $\forall t\in[0,q]$,

$$\begin{aligned}\left\|\mathbf{y}^{t+1,s}-\mathbf{J}\mathbf{y}^{t+1,s}\right\|^2 &\leq \frac{1+\lambda^2}{2\lambda^2}\left\|\mathbf{W}\mathbf{y}^{t,s}-\mathbf{J}\mathbf{y}^{t,s}\right\|^2 + \frac{1+\lambda^2}{1-\lambda^2}\left\|\mathbf{v}^{t,s}-\mathbf{v}^{t-1,s}\right\|^2\\ &\leq \frac{1+\lambda^2}{2}\left\|\mathbf{y}^{t,s}-\mathbf{J}\mathbf{y}^{t,s}\right\|^2 + \frac{2}{1-\lambda^2}\left\|\mathbf{v}^{t,s}-\mathbf{v}^{t-1,s}\right\|^2, \end{aligned} \tag{3.38}$$

where the last line is due to Lemma 3.2.4. Next, we derive upper bounds for the last term in (3.38) under different ranges of t and s.

(ii) $\forall t\in[1,q]$ **and** $\forall s\geq 1$**.** By the update of each local $\mathbf{v}_i^{t,s}$, we have that

$$\begin{aligned}\mathbb{E}\Big[\left\|\mathbf{v}^{t,s}-\mathbf{v}^{t-1,s}\right\|^2|\mathcal{F}^{t,s}\Big] &= \sum_{i=1}^n \mathbb{E}\left[\left\|\frac{1}{B}\sum_{l=1}^B\Big(\nabla f_{i,\tau_{i,l}^{t,s}}(\mathbf{x}_i^{t,s})-\nabla f_{i,\tau_{i,l}^{t,s}}(\mathbf{x}_i^{t-1,s})\Big)\right\|^2\Big|\mathcal{F}^{t,s}\right]\\ &\leq \frac{1}{B}\sum_{i=1}^n\sum_{l=1}^B \mathbb{E}\left[\left\|\nabla f_{i,\tau_{i,l}^{t,s}}(\mathbf{x}_i^{t,s})-\nabla f_{i,\tau_{i,l}^{t,s}}(\mathbf{x}_i^{t-1,s})\right\|^2\Big|\mathcal{F}^{t,s}\right]\\ &\leq L^2\left\|\mathbf{x}^{t,s}-\mathbf{x}^{t-1,s}\right\|^2. \end{aligned} \tag{3.39}$$

where the last line is due to Lemma 3.2.9. To proceed, we further use (3.33) and (3.17) to refine (3.39) as follows: $\forall s\geq 1$ and $\forall t\in[1,q]$,

$$\begin{aligned}\mathbb{E}\Big[\left\|\mathbf{v}^{t,s}-\mathbf{v}^{t-1,s}\right\|^2|\mathcal{F}^{t,s}\Big] &\leq 3L^2\left\|\mathbf{x}^{t,s}-\mathbf{J}\mathbf{x}^{t,s}\right\|^2 + 3n\alpha^2L^2\left\|\overline{\mathbf{v}}^{t-1,s}\right\|^2 + 3L^2\left\|\mathbf{x}^{t-1,s}-\mathbf{J}\mathbf{x}^{t-1,s}\right\|^2\\ &\leq 3n\alpha^2L^2\left\|\overline{\mathbf{v}}^{t-1,s}\right\|^2 + 9L^2\left\|\mathbf{x}^{t-1,s}-\mathbf{J}\mathbf{x}^{t-1,s}\right\|^2 + 6\alpha^2L^2\left\|\mathbf{y}^{t,s}-\mathbf{J}\mathbf{y}^{t,s}\right\|^2. \end{aligned} \tag{3.40}$$

We take the expectation of (3.40) and use it in (3.38) to obtain: $\forall s\geq 1$ and $\forall t\in[1,q]$,

$$\begin{aligned}\mathbb{E}\Big[\left\|\mathbf{y}^{t+1,s}-\mathbf{J}\mathbf{y}^{t+1,s}\right\|^2\Big] \leq& \left(\frac{1+\lambda^2}{2}+\frac{12\alpha^2L^2}{1-\lambda^2}\right)\mathbb{E}\Big[\left\|\mathbf{y}^{t,s}-\mathbf{J}\mathbf{y}^{t,s}\right\|^2\Big]\\ &+\frac{18L^2}{1-\lambda^2}\mathbb{E}\Big[\left\|\mathbf{x}^{t-1,s}-\mathbf{J}\mathbf{x}^{t-1,s}\right\|^2\Big] + \frac{6n\alpha^2L^2}{1-\lambda^2}\mathbb{E}\Big[\left\|\overline{\mathbf{v}}^{t-1,s}\right\|^2\Big].\end{aligned}$$

The second statement in the lemma follows by the fact that $\frac{1+\lambda^2}{2}+\frac{12\alpha^2L^2}{1-\lambda^2}\leq\frac{3+\lambda^2}{4}$ if $0<\alpha\leq\frac{1-\lambda^2}{4\sqrt{3}L}$.

(iii) $t=0$ **and** $\forall s\geq 2$**.** By the update of `GT-SARAH`, we observe that: $\forall s\geq 2$,

$$\begin{aligned}\left\|\mathbf{v}^{0,s}-\mathbf{v}^{-1,s}\right\|^2 &= \left\|\nabla\mathbf{f}(\mathbf{x}^{q+1,s-1})-\mathbf{v}^{q,s-1}\right\|^2\\ &= \left\|\nabla\mathbf{f}(\mathbf{x}^{q+1,s-1})-\nabla\mathbf{f}(\mathbf{x}^{q,s-1})+\nabla\mathbf{f}(\mathbf{x}^{q,s-1})-\mathbf{v}^{q,s-1}\right\|^2\\ &\leq 2L^2\left\|\mathbf{x}^{q+1,s-1}-\mathbf{x}^{q,s-1}\right\|^2 + 2\left\|\nabla\mathbf{f}\big(\mathbf{x}^{q,s-1}\big)-\mathbf{v}^{q,s-1}\right\|^2,\\ &\leq 6L^2\left\|\mathbf{x}^{q+1,s-1}-\mathbf{J}\mathbf{x}^{q+1,s-1}\right\|^2 + 6n\alpha^2L^2\left\|\overline{\mathbf{v}}^{q,s-1}\right\|^2\\ &\quad + 6L^2\left\|\mathbf{x}^{q,s-1}-\mathbf{J}\mathbf{x}^{q,s-1}\right\|^2 + 2\left\|\nabla\mathbf{f}\big(\mathbf{x}^{q,s-1}\big)-\mathbf{v}^{q,s-1}\right\|^2\\ &\leq 18L^2\left\|\mathbf{x}^{q,s-1}-\mathbf{J}\mathbf{x}^{q,s-1}\right\|^2 + 6n\alpha^2L^2\left\|\overline{\mathbf{v}}^{q,s-1}\right\|^2\\ &\quad + 12\alpha^2L^2\left\|\mathbf{y}^{q+1,s-1}-\mathbf{J}\mathbf{y}^{q+1,s-1}\right\|^2 + 2\left\|\nabla\mathbf{f}\big(\mathbf{x}^{q,s-1}\big)-\mathbf{v}^{q,s-1}\right\|^2, \end{aligned} \tag{3.41}$$

where the first inequality uses the L-smoothness of each f_i, the second inequality uses (3.33), and the last inequality uses (3.17). Taking the expectation of (3.41) and then using Lemma 3.2.11 gives: $\forall s \geq 2$,

$$\begin{aligned}\mathbb{E}\Big[\big\|\mathbf{v}^{0,s}-\mathbf{v}^{-1,s}\big\|^2\Big] \leq\ & 18L^2\mathbb{E}\Big[\big\|\mathbf{x}^{q,s-1}-\mathbf{J}\mathbf{x}^{q,s-1}\big\|^2\Big] + 6n\alpha^2L^2\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s-1}\big\|^2\Big] \\ & + 12\alpha^2L^2\mathbb{E}\Big[\big\|\mathbf{y}^{q+1,s-1}-\mathbf{J}\mathbf{y}^{q+1,s-1}\big\|^2\Big] \\ & + \frac{12L^2}{B}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\mathbf{x}^{t,s-1}-\mathbf{J}\mathbf{x}^{t,s-1}\big\|^2\Big]. \end{aligned} \tag{3.42}$$

We recall from (3.38) that $\forall s \geq 2$,

$$\big\|\mathbf{y}^{1,s}-\mathbf{J}\mathbf{y}^{1,s}\big\|^2 \leq \frac{1+\lambda^2}{2}\big\|\mathbf{y}^{q+1,s-1}-\mathbf{J}\mathbf{y}^{q+1,s-1}\big\|^2 + \frac{2}{1-\lambda^2}\big\|\mathbf{v}^{0,s}-\mathbf{v}^{-1,s}\big\|^2. \tag{3.43}$$

We finally apply (3.42) to (3.43) to obtain: $\forall s \geq 2$,

$$\begin{aligned}\mathbb{E}\Big[\big\|\mathbf{y}^{1,s}-\mathbf{J}\mathbf{y}^{1,s}\big\|^2\Big] \leq\ & \left(\frac{1+\lambda^2}{2}+\frac{24\alpha^2L^2}{1-\lambda^2}\right)\mathbb{E}\Big[\big\|\mathbf{y}^{q+1,s-1}-\mathbf{J}\mathbf{y}^{q+1,s-1}\big\|^2\Big] \\ & + \frac{36L^2}{1-\lambda^2}\mathbb{E}\Big[\big\|\mathbf{x}^{q,s-1}-\mathbf{J}\mathbf{x}^{q,s-1}\big\|^2\Big] + \frac{12n\alpha^2L^2}{1-\lambda^2}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s-1}\big\|^2\Big] \\ & + \frac{24L^2}{B(1-\lambda^2)}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\mathbf{x}^{t,s-1}-\mathbf{J}\mathbf{x}^{t,s-1}\big\|^2\Big]. \end{aligned}$$

We note that $\frac{1+\lambda^2}{2}+\frac{24\alpha^2L^2}{1-\lambda^2} \leq \frac{3+\lambda^2}{4}$ if $0<\alpha\leq\frac{1-\lambda^2}{4\sqrt{6}L}$ and then the third statement in the lemma follows. □

With the help of (3.16) and Lemma 3.2.12, we now abstract `GT-SARAH` with an LTI system to quantify jointly the state consensus and the gradient tracking error.

Lemma 3.2.13. *If the step-size α follows that $0<\alpha\leq\frac{1-\lambda^2}{4\sqrt{6}L}$, then we have*

$$\mathbf{u}^{t,s} \leq \mathbf{G}\mathbf{u}^{t-1,s} + \mathbf{b}^{t-1,s}, \qquad \forall s\in[1,S] \text{ and } t\in[1,q], \tag{3.44}$$

$$\mathbf{u}^{0,s} \leq \mathbf{G}\mathbf{u}^{q,s-1} + \mathbf{b}^{q,s-1} + \sum_{t=0}^{q}\Big(\mathbf{b}^{t,s-1}+\mathbf{H}\mathbf{u}^{t,s-1}\Big), \qquad \forall s\in[2,S], \tag{3.45}$$

where, $\forall s\geq 1$ and $\forall t\in[0,q]$,

$$\mathbf{u}^{t,s} := \begin{bmatrix} \frac{1}{n}\mathbb{E}\Big[\big\|\mathbf{x}^{t,s}-\mathbf{J}\mathbf{x}^{t,s}\big\|^2\Big] \\ \frac{1}{nL^2}\mathbb{E}\Big[\big\|\mathbf{y}^{t+1,s}-\mathbf{J}\mathbf{y}^{t+1,s}\big\|^2\Big]\end{bmatrix}, \qquad \mathbf{b} := \begin{bmatrix} 0 \\ \frac{12\alpha^2}{1-\lambda^2}\end{bmatrix}, \qquad \mathbf{b}^{t,s} := \mathbf{b}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s}\big\|^2\Big],$$

$$\mathbf{G} := \begin{bmatrix} \frac{1+\lambda^2}{2} & \frac{2\alpha^2L^2}{1-\lambda^2} \\ \frac{18}{1-\lambda^2} & \frac{3+\lambda^2}{4}\end{bmatrix}, \qquad \mathbf{H} := \begin{bmatrix} 0 & 0 \\ \frac{42}{1-\lambda^2} & 0\end{bmatrix}.$$

Proof. Write the inequalities in (3.16) and Lemma 3.2.12 jointly in a matrix form. □

We next derive the range of the step-size α such that $\rho(\mathbf{G}) < 1$, i.e. the LTI system does not diverge, with the help of the following lemma.

Lemma 3.2.14 ([36]). *Let $\mathbf{X} \in \mathbb{R}^{d\times d}$ be (entry-wise) non-negative and $\mathbf{x} \in \mathbb{R}^d$ be (entry-wise) positive. If $\mathbf{X}\mathbf{x} < \mathbf{x}$ (entry-wise), then $\rho(\mathbf{X}) < 1$.*

Lemma 3.2.15. *If the step-size α follows that $0 < \alpha < \frac{(1-\lambda^2)^2}{8\sqrt{5}L}$, then $\rho(\mathbf{G}) < 1$ and therefore $\sum_{k=0}^{\infty}\mathbf{G}^k$ is convergent such that $\sum_{k=0}^{\infty}\mathbf{G}^k = (\mathbf{I}_2 - \mathbf{G})^{-1}$.*

Proof. In the light of Lemma 3.2.14, we solve the range of α and a positive vector $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2]^\top$ such that $\mathbf{G}\boldsymbol{\varepsilon} < \boldsymbol{\varepsilon}$, which is equivalent to the following two inequalities.

$$\begin{cases} \dfrac{1+\lambda^2}{2}\varepsilon_1 + \dfrac{2\alpha^2L^2}{1-\lambda^2}\varepsilon_2 < \varepsilon_1 \\ \dfrac{18}{1-\lambda^2}\varepsilon_1 + \dfrac{3+\lambda^2}{4}\varepsilon_2 < \varepsilon_2 \end{cases} \iff \begin{cases} \alpha^2 < \dfrac{(1-\lambda^2)^2}{4L^2}\dfrac{\varepsilon_1}{\varepsilon_2} \\ \dfrac{\varepsilon_1}{\varepsilon_2} < \dfrac{(1-\lambda^2)^2}{72} \end{cases} \tag{3.46}$$

According to the second inequality of (3.46), we set $\varepsilon_1/\varepsilon_2 = (1-\lambda^2)^2/80$ and the proof follows by using it in the first inequality of (3.46) to solve for the range of α. □

Based on Lemma 3.2.15, the LTI system is stable under an appropriate step-size α and therefore we can solve the LTI system to obtain the following lemma, the proof of which is deferred to Section 3.2.7.7 for the ease of exposition.

Lemma 3.2.16. *If $0 < \alpha < \frac{(1-\lambda^2)^2}{8\sqrt{5}L}$, then the following inequality holds.*

$$\Big(\mathbf{I}_2 - (\mathbf{I}_2 - \mathbf{G})^{-1}\mathbf{H}\Big)\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbf{u}^{t,s} \leq (\mathbf{I}_2 - \mathbf{G})^{-1}\mathbf{u}^{0,1} + 2(\mathbf{I}_2 - \mathbf{G})^{-1}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbf{b}^{t,s}.$$

Proof. See Section 3.2.7.7. □

In the following lemma, we compute $(\mathbf{I}_2 - \mathbf{G})^{-1}$ and $(\mathbf{I}_2 - \mathbf{G})^{-1}\mathbf{b}$.

Lemma 3.2.17. *If $0 < \alpha \leq \frac{(1-\lambda^2)^2}{24L}$, then the following entry-wise inequality holds:*

$$(\mathbf{I}_2 - \mathbf{G})^{-1} \leq \begin{bmatrix} \dfrac{4}{1-\lambda^2} & \dfrac{32\alpha^2L^2}{(1-\lambda^2)^3} \\ \dfrac{288}{(1-\lambda^2)^3} & \dfrac{8}{1-\lambda^2} \end{bmatrix}, \quad (\mathbf{I}_2 - \mathbf{G})^{-1}\mathbf{b} \leq \begin{bmatrix} \dfrac{384\alpha^4L^2}{(1-\lambda^2)^4} \\ \dfrac{96\alpha^2}{(1-\lambda^2)^2} \end{bmatrix}.$$

Proof. We first derive a lower bound for $\det(\mathbf{I}_2 - \mathbf{G})$. Note that if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{24L}$, then $\det(\mathbf{I}_2 - \mathbf{G}) = \frac{(1-\lambda^2)^2}{8} - \frac{36\alpha^2L^2}{(1-\lambda^2)^2} \geq \frac{(1-\lambda^2)^2}{16}$ and therefore

$$(\mathbf{I}_2 - \mathbf{G})^{-1} \leq \frac{16}{(1-\lambda^2)^2}\begin{bmatrix} \dfrac{1-\lambda^2}{4} & \dfrac{2\alpha^2L^2}{1-\lambda^2} \\ \dfrac{18}{1-\lambda^2} & \dfrac{1-\lambda^2}{2} \end{bmatrix} = \begin{bmatrix} \dfrac{4}{1-\lambda^2} & \dfrac{32\alpha^2L^2}{(1-\lambda^2)^3} \\ \dfrac{288}{(1-\lambda^2)^3} & \dfrac{8}{1-\lambda^2} \end{bmatrix},$$

and the proof follows by the definition of $\mathbf{b}$ in Lemma 3.2.13. □

Using Lemma 3.2.17, we have: if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{8\sqrt{42}L}$,

$$\mathbf{I}_2 - (\mathbf{I}_2 - \mathbf{G})^{-1}\mathbf{H} \geq \begin{bmatrix} 1 - \frac{1344\alpha^2 L^2}{(1-\lambda^2)^4} & 0 \\ -\frac{336}{(1-\lambda^2)^2} & 1 \end{bmatrix} \geq \begin{bmatrix} \frac{1}{2} & 0 \\ -\frac{336}{(1-\lambda^2)^2} & 1 \end{bmatrix}. \tag{3.47}$$

Finally, we apply (3.47) and Lemma 3.2.17 to Lemma 3.2.16 to obtain

$$\frac{1}{2n}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\big\|^2\Big] \leq \frac{32\alpha^2}{n(1-\lambda^2)^3}\mathbb{E}\Big[\big\|\mathbf{y}^{1,1} - \mathbf{J}\mathbf{y}^{1,1}\big\|^2\Big] + \frac{768\alpha^4 L^2}{(1-\lambda^2)^4}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s}\big\|^2\Big]. \tag{3.48}$$

The proof of Lemma 3.2.7 follows by applying the first statement in Lemma 3.2.12 to (3.48).

3.2.7.5 Proof of Lemma 3.2.8

We have: $\forall s \geq 1$ and $\forall t \in [0, q]$,

$$\begin{aligned}\frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\nabla F(\mathbf{x}_i^{t,s})\big\|^2\Big] &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\nabla F(\mathbf{x}_i^{t,s}) - \nabla F(\overline{\mathbf{x}}^{t,s})\big\|^2\Big] + \mathbb{E}\Big[\big\|\nabla F(\overline{\mathbf{x}}^{t,s})\big\|^2\Big] \\ &\leq \frac{L^2}{n}\mathbb{E}\Big[\big\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\big\|^2\Big] + \mathbb{E}\Big[\big\|\nabla F(\overline{\mathbf{x}}^{t,s})\big\|^2\Big],\end{aligned} \tag{3.49}$$

where the second line is due to the L-smoothness of F. Since F is bounded below by F^*, we may apply (3.49) to Lemma 3.2.5 to obtain the following: if $0 < \alpha \leq \frac{1}{2L}$,

$$\begin{aligned}F^* \leq F(\overline{\mathbf{x}}^{0,1}) &- \frac{\alpha}{4n}\sum_{i=1}^{n}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\big\|\nabla F(\mathbf{x}_i^{t,s})\big\|^2\right] - \frac{\alpha}{4}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\big\|\overline{\mathbf{v}}^{t,s}\big\|^2\right] \\ &+ \alpha\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\big\|\overline{\mathbf{v}}^{t,s} - \overline{\nabla\mathbf{f}}(\mathbf{x}^{t,s})\big\|^2\right] + \frac{3\alpha L^2}{2}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\frac{\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\|^2}{n}\Big].\end{aligned} \tag{3.50}$$

We then apply Lemma 3.2.6 to (3.50) to obtain: if $0 < \alpha \leq \frac{1}{2L}$,

$$\begin{aligned}F^* \leq F(\overline{\mathbf{x}}^{0,1}) &- \frac{\alpha}{4n}\sum_{i=1}^{n}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\big\|\nabla F(\mathbf{x}_i^{t,s})\big\|^2\right] - \frac{\alpha}{8}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\big\|\overline{\mathbf{v}}^{t,s}\big\|^2\right] \\ &+ \alpha L^2\left(\frac{3}{2} + \frac{6q}{nB}\right)\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\frac{\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\|^2}{n}\Big] - \frac{\alpha}{8}\left(1 - \frac{24q\alpha^2 L^2}{nB}\right)\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\left[\big\|\overline{\mathbf{v}}^{t,s}\big\|^2\right].\end{aligned} \tag{3.51}$$

If $0 < \alpha \leq \frac{\sqrt{nB}}{2\sqrt{6q}L}$ then $1 - \frac{24\alpha^2 qL^2}{nB} \geq 0$ and thus the last term in (3.51) may be dropped. We finally apply Lemma 3.2.7 to (3.51) to obtain: if $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{4\sqrt{42}}, \sqrt{\frac{nB}{6q}}\right\}\frac{1}{2L}$,

$$\begin{aligned}F^* \leq F(\overline{\mathbf{x}}^{0,1}) &- \frac{\alpha}{4n}\sum_{i=1}^{n}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\nabla F(\mathbf{x}_i^{t,s})\big\|^2\Big] + \left(\frac{7}{4} + \frac{6q}{nB}\right)\frac{64\alpha^3 L^2}{(1-\lambda^2)^3}\frac{\|\nabla\mathbf{f}(\mathbf{x}^{0,1})\|^2}{n} \\ &- \frac{\alpha L^2}{4}\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\frac{\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\|^2}{n}\Big] - \frac{\alpha}{8}\left(1 - \left(\frac{7}{4} + \frac{6q}{nB}\right)\frac{12288\alpha^4 L^4}{(1-\lambda^2)^4}\right)\sum_{s=1}^{S}\sum_{t=0}^{q}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t,s}\big\|^2\Big].\end{aligned}$$

We observe that if $0 < \alpha \leq \left(\frac{4nB}{7nB+24q}\right)^{1/4}\frac{1-\lambda^2}{12L}$, then $1 - (\frac{7}{4} + \frac{6q}{nB})\frac{12288\alpha^4 L^4}{(1-\lambda^2)^4} \geq 0$ and thus the last term in the above inequality may be dropped; the proof follows.

3.2.7.6 Proof of Lemma 3.2.11

In the following, we use the notation in (3.25). Using the update of each local recursive gradient estimator $\mathbf{v}_i^{t,s}$, we have that: $\forall i \in \mathcal{V}, \forall s \geq 1$ and $t \in [1, q]$,

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\mathbf{v}_i^{t,s} - \nabla f_i(\mathbf{x}_i^{t,s})\big\|^2 \big| \mathcal{F}^{t,s}\Big] \\
&= \mathbb{E}\Big[\Big\|\widehat{\nabla}_i^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big) + \mathbf{v}_i^{t-1,s} - \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\Big\|^2 \Big| \mathcal{F}^{t,s}\Big] \\
&= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{l=1}^{B}\Big(\widehat{\nabla}_{i,l}^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\Big)\right\|^2 \Big| \mathcal{F}^{t,s}\right] + \big\|\mathbf{v}_i^{t-1,s} - \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\big\|^2, \\
&= \frac{1}{B^2}\sum_{l=1}^{B}\mathbb{E}\Big[\Big\|\widehat{\nabla}_{i,l}^{t,s} - \nabla f_i\big(\mathbf{x}_i^{t,s}\big) + \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\Big\|^2 \Big| \mathcal{F}^{t,s}\Big] + \big\|\mathbf{v}_i^{t-1,s} - \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\big\|^2, \\
&\leq \frac{1}{B^2}\sum_{l=1}^{B}\mathbb{E}\Big[\Big\|\nabla f_{i,\tau_{i,l}^{t,s}}\big(\mathbf{x}_i^{t,s}\big) - \nabla f_{i,\tau_{i,l}^{t,s}}\big(\mathbf{x}_i^{t-1,s}\big)\Big\|^2 \Big| \mathcal{F}^{t,s}\Big] + \big\|\mathbf{v}_i^{t-1,s} - \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\big\|^2, \\
&\leq \frac{L^2}{B}\big\|\mathbf{x}_i^{t,s} - \mathbf{x}_i^{t-1,s}\big\|^2 + \big\|\mathbf{v}_i^{t-1,s} - \nabla f_i\big(\mathbf{x}_i^{t-1,s}\big)\big\|^2.
\end{aligned}
$$

The above derivations follow a similar line of arguments as in the proof of Lemma 3.2.10 and hence we omit the details here. Summing up the last inequality above over i from 1 to n and taking the expectation, we have: $\forall s \geq 1$ and $t \in [1, q]$,

$$
\mathbb{E}\Big[\big\|\mathbf{v}^{t,s} - \nabla\mathbf{f}\big(\mathbf{x}^{t,s}\big)\big\|^2\Big] \leq \mathbb{E}\left[\frac{L^2}{B}\big\|\mathbf{x}^{t,s} - \mathbf{x}^{t-1,s}\big\|^2 + \big\|\mathbf{v}^{t-1,s} - \nabla\mathbf{f}\big(\mathbf{x}^{t-1,s}\big)\big\|^2\right]. \tag{3.52}
$$

Recall from (3.33) that $\forall s \geq 1$ and $t \in [1, q]$,

$$
\big\|\mathbf{x}^{t,s} - \mathbf{x}^{t-1,s}\big\|^2 \leq 3\big\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\big\|^2 + 3n\alpha^2\big\|\overline{\mathbf{v}}^{t-1,s}\big\|^2 + 3\big\|\mathbf{x}^{t-1,s} - \mathbf{J}\mathbf{x}^{t-1,s}\big\|^2. \tag{3.53}
$$

Applying (3.53) to (3.52) obtains: $\forall s \geq 1$ and $t \in [1, q]$,

$$
\begin{aligned}
\mathbb{E}\Big[\big\|\mathbf{v}^{t,s} - \nabla\mathbf{f}(\mathbf{x}^{t,s})\big\|^2\Big] \leq{}& \mathbb{E}\Big[\big\|\mathbf{v}^{t-1,s} - \nabla\mathbf{f}(\mathbf{x}^{t-1,s})\big\|^2\Big] + \frac{3n\alpha^2L^2}{B}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{t-1,s}\big\|^2\Big] \\
&+ \frac{3L^2}{B}\mathbb{E}\Big[\big\|\mathbf{x}^{t,s} - \mathbf{J}\mathbf{x}^{t,s}\big\|^2\Big] + \frac{3L^2}{B}\mathbb{E}\Big[\big\|\mathbf{x}^{t-1,s} - \mathbf{J}\mathbf{x}^{t-1,s}\big\|^2\Big].
\end{aligned} \tag{3.54}
$$

Recall that $\mathbf{v}^{0,s} = \nabla\mathbf{f}(\mathbf{x}^{0,s}), \forall s \geq 1$, and we take the telescoping sum of (3.54) over t to obtain: $\forall s \geq 1$ and $t \in [1, q]$,

$$
\begin{aligned}
\mathbb{E}\Big[\big\|\mathbf{v}^{t,s} - \nabla\mathbf{f}(\mathbf{x}^{t,s})\big\|^2\Big] \leq{}& \frac{3n\alpha^2L^2}{B}\sum_{u=1}^{t}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}^{u-1,s}\big\|^2\Big] + \frac{3L^2}{B}\sum_{u=1}^{t}\mathbb{E}\Big[\big\|\mathbf{x}^{u,s} - \mathbf{J}\mathbf{x}^{u,s}\big\|^2\Big] \\
&+ \frac{3L^2}{B}\sum_{u=1}^{t}\mathbb{E}\Big[\big\|\mathbf{x}^{u-1,s} - \mathbf{J}\mathbf{x}^{u-1,s}\big\|^2\Big].
\end{aligned}
$$

The proof follows by merging the last two terms on the RHS of the inequality above.

Table 3.3: The one-on-one mapping between the single-loop sequences $\{\mathbf{u}^k\}, \{\mathbf{b}^k\}$ for $k \in [0, S(q+1)-1]$ and the double-loop sequences $\{\mathbf{u}^{t,s}\}, \{\mathbf{b}^{t,s}\}$ for $s \in [1, S]$ and $t \in [0, q]$.

k	(t, s)
$0, \cdots, q$	$(0,1), \cdots, (q,1)$
$q+1, \cdots, 2q+1$	$(0,2), \cdots, (q,2)$
$\cdots$	$\cdots$
$(S-1)(q+1), \cdots, S(q+1)-1$	$(0,S), \cdots, (q,S)$

3.2.7.7 Proof of Lemma 3.2.16

Step 1: A loop-less dynamical system. For the ease of calculations, we first write the LTI system in Lemma 3.2.13 in a equivalent *loopless* form. To do this, we unroll the original *double loop* sequences $\{\mathbf{u}^{t,s}\}$ and $\{\mathbf{b}^{t,s}\}$, where $t \in [0, q]$ and $s \in [1, S]$, respectively as *loopless* sequences $\{\mathbf{u}^k\}$ and $\{\mathbf{b}^k\}$, where $k \in [0, S(q+1)-1]$, as follows:

$$\mathbf{u}^k := \mathbf{u}^{t,s}, \quad \mathbf{b}^k := \mathbf{b}^{t,s}, \qquad \text{where } k = t + (s-1)(q+1), \tag{3.55}$$

for $t \in [0, q]$ and $s \in [1, S]$. Reversely, given $\mathbf{u}^k$ and $\mathbf{b}^k$, for $k \in [0, S(q+1)-1]$, we can find their positions in the original double loop sequence, $\mathbf{u}^{t,s}$ and $\mathbf{b}^{t,s}$, by

$$t = \text{ mod } (k, q+1) \text{ and } s = \lfloor k/(q+1) \rfloor + 1, \qquad \text{for } k \in [0, S(q+1)-1]. \tag{3.56}$$

This one-on-one correspondence is visualized in Table 3.2.7.7.

With (3.55) and (3.56) at hand, it can be verified that the following single-loop system is equivalent to the double loop system in (3.44) and (3.45). For $k \in [1, S(q+1)-1]$,

$$\mathbf{u}^k \leq \mathbf{G}\mathbf{u}^{k-1} + \mathbf{b}^{k-1}, \qquad \text{if} \quad \text{mod } (k, q+1) \neq 0. \tag{3.57}$$

$$\mathbf{u}^{z(q+1)} \leq \mathbf{G}\mathbf{u}^{z(q+1)-1} + \mathbf{b}^{z(q+1)-1} + \sum_{r=(z-1)(q+1)}^{z(q+1)-1} \mathbf{h}^r, \qquad \forall z \in [1, S-1], \tag{3.58}$$

where

$$\mathbf{h}^k := \mathbf{b}^k + \mathbf{H}\mathbf{u}^k.$$

The system in (3.57) and (3.58) can be further written equivalently as the following: $\forall k \in [1, S(q+1)-1]$,

$$\mathbf{u}^k \leq \mathbf{G}\mathbf{u}^{k-1} + \mathbf{d}^k, \tag{3.59}$$

where

$$\mathbf{d}^k := \mathbf{b}^{k-1} + \mathbb{1}\{\text{mod } (k, q+1) = 0\} \sum_{r=k-(q+1)}^{k-1} \mathbf{h}^r,$$

such that $\mathbb{1}\{\cdot\}$ is the indicator function of an event, and $\sum_{r=k-(q+1)}^{k-1} \mathbf{h}^r := 0$ for $k \in [1, q]$.

Step 2: Analyzing the recursion. We recursively apply (3.59) over k to obtain: $\forall k \in [1, S(q+1)-1]$,

$$\mathbf{u}^k \leq \mathbf{G}^k \mathbf{u}^0 + \sum_{r=1}^{k} \mathbf{G}^{k-r} \mathbf{d}^r. \tag{3.60}$$

Summing up (3.60) over k from 0 to $S(q+1)-1$ gives: if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{8\sqrt{5}L}$,

$$\begin{aligned}
\sum_{k=0}^{S(q+1)-1} \mathbf{u}^k &\leq \sum_{k=0}^{S(q+1)-1} \mathbf{G}^k \mathbf{u}^0 + \sum_{k=1}^{S(q+1)-1} \sum_{r=1}^{k} \mathbf{G}^{k-r} \mathbf{d}^r \\
&\leq \left(\sum_{k=0}^{\infty} \mathbf{G}^k \right) \mathbf{u}^0 + \sum_{k=1}^{S(q+1)-1} \left(\sum_{r=1}^{\infty} \mathbf{G}^r \right) \mathbf{d}^k \\
&= (\mathbf{I}_2 - \mathbf{G})^{-1} \mathbf{u}^0 + (\mathbf{I}_2 - \mathbf{G})^{-1} \sum_{k=1}^{S(q+1)-1} \mathbf{d}^k.
\end{aligned} \tag{3.61}$$

To proceed, we recall the definition of $\mathbf{d}^k$ and $\mathbf{h}^k$ and observe that

$$\begin{aligned}
\sum_{k=1}^{S(q+1)-1} \mathbf{d}^k &= \sum_{k=0}^{S(q+1)-2} \mathbf{b}^k + \sum_{k=1}^{S(q+1)-1} \left(\mathbb{1}\{\text{mod } (k, q+1) = 0\} \sum_{r=k-(q+1)}^{k-1} \mathbf{h}^r \right) \\
&= \sum_{k=0}^{S(q+1)-2} \mathbf{b}^k + \sum_{z=1}^{S-1} \left(\sum_{r=(z-1)(q+1)}^{z(q+1)-1} \mathbf{h}^r \right) \\
&= \sum_{k=0}^{S(q+1)-2} \mathbf{b}^k + \sum_{k=0}^{(S-1)(q+1)-1} \mathbf{h}^k \\
&\leq 2 \sum_{k=0}^{S(q+1)-1} \mathbf{b}^k + \sum_{k=0}^{(S-1)(q+1)-1} \mathbf{H}\mathbf{u}^k,
\end{aligned} \tag{3.62}$$

where the first line and the last line are due to the definition of $\mathbf{d}^k$ and $\mathbf{h}^k$ respectively. Finally, we use (3.62) in (3.61) to obtain: if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{8\sqrt{5}L}$, then

$$\sum_{k=0}^{S(q+1)-1} \mathbf{u}^k \leq (\mathbf{I}_2 - \mathbf{G})^{-1} \mathbf{u}^0 + 2(\mathbf{I}_2 - \mathbf{G})^{-1} \sum_{k=0}^{S(q+1)-1} \mathbf{b}^k + (\mathbf{I}_2 - \mathbf{G})^{-1} \mathbf{H} \sum_{k=0}^{S(q+1)-1} \mathbf{u}^k,$$

which is the same as

$$\left(\mathbf{I}_2 - (\mathbf{I}_2 - \mathbf{G})^{-1} \mathbf{H}\right) \sum_{k=0}^{S(q+1)-1} \mathbf{u}^k \leq (\mathbf{I}_2 - \mathbf{G})^{-1} \mathbf{u}^0 + 2(\mathbf{I}_2 - \mathbf{G})^{-1} \sum_{k=0}^{S(q+1)-1} \mathbf{b}^k.$$

We conclude the proof of Lemma 3.2.16 by rewriting the above inequality in the original double loop form.

3.3 Stochastic incremental variance reduction

In this section, we revisit and analyze the **GT-SAGA** algorithm, originally proposed in Chapter 2 under strong convexity, for solving the decentralized smooth non-convex finite-sum problem (3.1).

3.3.1 Main contributions

We analyze **GT-SAGA**, a *single-timescale randomized incremental* gradient method, originally proposed in [24] for strongly-convex problems, and show that it achieves fast convergence in non-convex settings. At the node level, **GT-SAGA** adopts a local SAGA-type [57, 60, 138–140] randomized incremental approach to obtain variance-reduced estimates of local batch gradients, by leveraging historical component gradient information. At the network level, **GT-SAGA** employs a gradient tracking mechanism [55, 65] to fuse the local batch gradient estimates, obtained from the local SAGA procedures, to track the global batch gradient. These are the two building blocks that amount to the fast convergence and robustness to heterogeneous data in **GT-SAGA** for non-convex problems. Compared with the existing two-timescale variance reduced methods [135, 141] for decentralized non-convex optimization, **GT-SAGA** is single-timescale and eliminates completely the need of batch gradient computations and periodic network synchronizations, and is hence much easier to implement especially in ad hoc settings; see Remarks 3.3.1 and 3.3.2 for further discussion. Our main technical contributions are summarized as follows:

- **General smooth non-convex problems.** For this problem class, we show the asymptotic convergence of **GT-SAGA** to a first-order stationary point in the almost sure and mean-squared sense. In a big-data regime, where the local batch size m is very large, **GT-SAGA** achieves a network topology-independent convergence rate, leading to a non-asymptotic linear speedup compared with the centralized SAGA [60] at a single node. In large-scale network regimes, i.e., when the number of the nodes and the network spectral gap inverse are relatively large compared to the local batch size m, we show that **GT-SAGA** outperforms the existing best known convergence rate [141]. We also introduce a measure of function heterogeneity across the nodes. Based on this measure, we show that the effect of function heterogeneity on the convergence rate of **GT-SAGA** appears in a fashion that is separable from the effects of local batch size and the network spectral gap. As a consequence, the effect of function heterogeneity often diminishes when the local batch size is large and/or the connectivity of the network is weak, demonstrating the robustness of **GT-SAGA** to function heterogeneity. In contrast, the state-of-the-art decentralized non-convex variance-reduced method [141] does not achieve such separation and hence has worse convergence rate than **GT-SAGA** when the function heterogeneity is large and the network is weakly connected. These improvements are achieved by leveraging the conditional unbiasedness of SAGA estimators to obtain tighter bounds in the stochastic gradient tracking analysis; see Remarks 3.3.3, 3.3.4, and 3.3.5.

- **Smooth non-convex problems under the Polyak-Lojasiewicz (PL) condition.** For this problem class, we show that **GT-SAGA** achieves linear convergence to an optimal solution in expectation. To

the best of our knowledge, this is the first linear rate result for decentralized variance-reduced methods under the PL condition, while the existing ones require strong convexity [22–24, 120, 142]. This generalization is non-trivial since the existing analysis essentially uses the unique optimal solution under strong convexity as a reference point to bound related error terms, while the PL condition allows for the existence of multiple optimal solutions. In comparison with the existing linearly-convergent, decentralized deterministic batch gradient methods under the PL condition [4, 143, 144], **GT-SAGA** provably achieves faster linear rate, in terms of the component gradient computation complexity at each node, when the local batch size m is large, demonstrating the advantage of the employed variance reduction technique. In a big-data regime where m is large enough, we show that the linear rate of **GT-SAGA** becomes network topology-independent. See Remarks 3.3.6 and 3.3.7 for details.

- **Technical analysis.** We note that our analysis of SAGA-type variance reduction procedures is different from the existing ones [60, 145], which require careful constructions of Lyapunov functions. We avoid such delicate constructions by adopting a direct analysis approach, based on linear time-invariant (LTI) dynamics, which may be of independent interest and perhaps more readily extendable to other non-convex problems. We note that the LTI dynamics-based analysis has mainly been used in convex problems in the existing literature of gradient tracking methods, e.g., [56, 67]. Somewhat surprisingly, a special case of our analysis, i.e., when the network is complete, provides the first linear rate result of the *original* centralized SAGA algorithm [57] under the PL condition. Indeed, the existing analysis [60, 145] is only applicable to a *modified* SAGA, which periodically restarts and samples its iterates; see Remark 3.3.8 for details. Our analysis is also substantially different from that of the existing decentralized non-convex variance-reduced methods [135, 141], where the variances of the stochastic gradients are bounded recursively, due to their hybrid nature. In contrast, we introduce a proper auxiliary sequence to bound the variance of **GT-SAGA**; see Subsection 3.3.5.2, 3.3.5.4 for details.

3.3.2 The non-convex GT-SAGA algorithm

GT-SAGA, built upon local SAGA estimators [57] and global gradient tracking [55, 65], is formally presented in Algorithm 4. We refer the readers to Chapter 2 for detailed discussion on the development of **GT-SAGA**. It should be noted, however, that Algorithm 4 is different from Algorithm 1 in terms of the sampling procedure, for ease of the convergence analysis. We require for conciseness that all nodes start at the same point, but the complexity results of **GT-SAGA** established here hold, up to factors of universal constants, for the case where the nodes are initialized differently. We comment on the practical implementation aspects of **GT-SAGA** in comparison with the existing approaches in the following remarks.

Algorithm 4 Non-convex **GT-SAGA** at each node i

Require: $\mathbf{x}_i^0 = \overline{\mathbf{x}}^0 \in \mathbb{R}^p$; $\alpha \in \mathbb{R}^+$; $\{\underline{w}_{ir}\}_{r=1}^n$; $\mathbf{z}_{i,j}^0 = \mathbf{x}_i^0, \forall j \in \{1, \cdots, m\}$; $\mathbf{y}_i^0 = \mathbf{0}_p$; $\mathbf{g}_i^{-1} = \mathbf{0}_p$.

for $k = 0, 1, 2, \cdots$ **do**

Select τ_i^k uniformly at random from $\{1, \cdots, m\}$;

Update the local stochastic gradient estimator:

$$\mathbf{g}_i^k = \nabla f_{i,\tau_i^k}(\mathbf{x}_i^k) - \nabla f_{i,\tau_i^k}(\mathbf{z}_{i,\tau_i^k}^k) + \frac{1}{m}\sum_{j=1}^m \nabla f_{i,j}(\mathbf{z}_{i,j}^k);$$

Update the local gradient tracker:

$$\mathbf{y}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\left(\mathbf{y}_r^k + \mathbf{g}_r^k - \mathbf{g}_r^{k-1}\right);$$

Update the local estimate of the solution:

$$\mathbf{x}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\left(\mathbf{x}_r^k - \alpha \mathbf{y}_r^{k+1}\right);$$

Select s_i^k uniformly at random from $\{1, \cdots, m\}$;

Set $\mathbf{z}_{i,j}^{k+1} = \mathbf{x}_i^k$ for $j = s_i^k$; $\mathbf{z}_{i,j}^{k+1} = \mathbf{z}_{i,j}^k$ for $j \neq s_i^k$;

end for

Remark 3.3.1 (**Single-timescale implementation**). The existing decentralized variance-reduced methods for non-convex optimization [135,141] are based on a two-timescale, double-loop implementation. Specifically, these methods, within each inner-loop, run a fixed number of stochastic gradient type iterations, while, at each outer-loop iteration, a local batch gradient is computed at each node. This double-loop nature imposes challenges on the practical implementation of the two methods in [135, 141]. First, periodic batch gradient computation incurs a synchronization overhead on the communication network and jeopardizes the actual wall-clock time when the networked nodes have largely heterogeneous computational capabilities. Second, these two methods have an additional parameter to tune, i.e., the length of each inner loop, other than the step-size. Although this parameter maybe be chosen as m [141], this particular choice may not lead to the best performance in practice. In sharp contrast, **GT-SAGA** admits a simple single-timescale implementation since it only evaluates *one* randomly selected component gradient at each iteration. Furthermore, it only has *one* parameter to tune, i.e., the step-size α. Therefore, **GT-SAGA** leads to significantly simpler implementation and tuning compared with the existing decentralized non-convex variance-reduced methods [135, 141], especially over large-scale ad-hoc networks. Finally, we note that **GT-SAGA** takes two successive communication rounds per iteration to transmit the state and gradient tracker respectively, as in

other gradient tracking-based methods, e.g., [4, 67, 135, 141].

Remark 3.3.2 (**Storage requirement**). To practically implement `GT-SAGA`, each node i needs to retain a gradient table $\{\nabla f_{i,j}(\mathbf{z}_{i,j}^k)\}_{j=1}^m$ of size $m \times p$ in general, which may be expensive. However, for certain structured problems, the size of the gradient table can be largely reduced [57]. For instance, in non-convex generalized linear models [146], each component function takes the form $f_{i,j}(\mathbf{x}) = \ell(\mathbf{x}^\top \boldsymbol{\theta}_{i,j})$, where $\ell : \mathbb{R} \to \mathbb{R}$ is a non-convex loss and $\boldsymbol{\theta}_{i,j}$ is the j-th data at the i-th node. Clearly, $\nabla f_{i,j}(\mathbf{x}) = \ell'(\mathbf{x}^\top \boldsymbol{\theta}_{i,j})\boldsymbol{\theta}_{i,j}$ and thus each node i only needs to retain $\{\ell'(\mathbf{z}_{i,j}^\top \boldsymbol{\theta}_{i,j})\}_{j=1}^m$, a gradient table of size $m \times 1$, since the data samples $\{\boldsymbol{\theta}_{i,j}\}_{j=1}^m$ are already stored locally. See Section 3.3.4.1 for numerical experiments based on one such example.

3.3.3 Main convergence results

We now enlist the assumptions of interest.

Assumption 3.3.1. *The family $\{\tau_i^k, s_i^k : i \in \mathcal{V}, k \geq 0\}$ of random variables in Algorithm 4 is independent.*

Assumption 3.3.1 is standard in stochastic gradient methods. Specifically, the index s_i^k used for updating the gradient table $\{\nabla f_{i,j}(\mathbf{z}_{i,j}^k)\}_{j=1}^m$ is sampled independently from the index τ_i^k used for updating the local SAGA estimator $\mathbf{g}_i^k$ per node per iteration. This independence requirement is straightforward to implement and is often posed to simplify the analysis of SAGA type estimators for non-convex problems [60, 145]; see Section 3.3.5.4 for analysis based on this assumption.

Assumption 3.3.2. *Each component function $f_{i,j} : \mathbb{R}^p \to \mathbb{R}$ is differentiable and L-smooth, i.e., there exists $L > 0$, such that $\|\nabla f_{i,j}(\mathbf{x}) - \nabla f_{i,j}(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, $\forall i \in \mathcal{V}$, $\forall j \in \{1, \cdots, m\}$. Moreover, the global function F is bounded below, i.e., $F^* := \inf_{\mathbf{x} \in \mathbb{R}^p} F(\mathbf{x}) > -\infty$.*

Under Assumption 3.3.2, the local batch functions $\{f_i\}_{i=1}^n$ and the global function F are L-smooth. We note that L stated in Assumption 3.3.2 is essentially the maximum of the smoothness parameters of all component functions. We further consider the case when *the global F* additionally satisfies the Polyak-Lojasiewicz (PL) condition described below.

Assumption 3.3.3. *The global function $F : \mathbb{R}^p \to \mathbb{R}$ satisfies $2\mu(F(\mathbf{x}) - F^*) \leq \|\nabla F(\mathbf{x})\|^2$, $\forall \mathbf{x} \in \mathbb{R}^p$, for some $\mu > 0$.*

The PL condition, originally introduced in [5], generalizes the notion of strong convexity to non-convex functions; see [147] for more discussion. When Assumption 3.3.3 holds, we denote $\kappa := \frac{L}{\mu} \geq 1$, which may be interpreted as the condition number of F. Note that the PL condition implies that every stationary point $\mathbf{x}^*$ of F, such that $\nabla F(\mathbf{x}^*) = \mathbf{0}_p$, is a global minimizer of F, while F is not necessarily convex.

Assumption 3.3.4. *The weight matrix* $\underline{\mathbf{W}} = \{\underline{w}_{ir}\} \in \mathbb{R}^{n\times n}$ *of the network is primitive and doubly-stochastic, i.e.,* $\underline{\mathbf{W}}\mathbf{1}_n = \mathbf{1}_n$, $\mathbf{1}_n^\top \underline{\mathbf{W}} = \mathbf{1}_n^\top$, *and* $\lambda := \lambda_2(\underline{\mathbf{W}}) \in [0,1)$, *where* $\lambda_2(\mathbf{W})$ *is the second largest singular value of* $\underline{\mathbf{W}}$.

Weight matrices that satisfy Assumption 3.3.4 may be designed for strongly-connected, weight-balanced, directed networks or for connected, undirected networks. We next discuss the performance metrics of **GT-SAGA** for different problem classes. For general smooth non-convex problems, we define the iteration complexity of **GT-SAGA** as the minimum number of iterations required to achieve an ϵ-accurate stationary point of the global function F, i.e.,

$$\inf\left\{K : \frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla F(\mathbf{x}_i^k)\|^2\big] \leq \epsilon\right\}.$$

When the global function F satisfies the PL condition, we define the iteration complexity of **GT-SAGA** as

$$\inf\left\{k : \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\big(F(\mathbf{x}_i^k) - F^*\big)\right] \leq \epsilon\right\}.$$

These are standard metrics for decentralized stochastic non-convex optimization methods [2,4,135,141]. We refer the iteration complexity as the the convergence rate metric of **GT-SAGA**, since it is the same as the communication and component gradient computation complexity at each node. We are now ready to state the main results of **GT-SAGA** in the next subsections and discuss their implications.

3.3.3.1 General smooth non-convex functions

In this section, we present the convergence results of **GT-SAGA** for general smooth non-convex functions.

Theorem 3.3.1. *Let Assumptions 3.3.1, 3.3.2, and 3.3.4 hold. If the step-size* α *of* ***GT-SAGA*** *satisfies* $0 < \alpha \leq \overline{\alpha}_1$, *where*

$$\overline{\alpha}_1 := \min\left\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{2n^{1/3}}{13m^{2/3}}, \frac{1}{2}, \frac{(1-\lambda^2)^{3/4}}{18\lambda^{1/2}m^{1/2}}\right\}\frac{1}{L},$$

then all nodes asymptotically agree on a stationary point in both mean-squared and almost sure sense, i.e., $\forall i, r \in \mathcal{V}$,

$$\mathbb{P}\Big(\lim_{k\to\infty}\|\mathbf{x}_i^k - \mathbf{x}_r^k\| = 0\Big) = 1, \quad \lim_{k\to\infty}\mathbb{E}\big[\|\mathbf{x}_i^k - \mathbf{x}_r^k\|^2\big] = 0,$$
$$\mathbb{P}\Big(\lim_{k\to\infty}\|\nabla F(\mathbf{x}_i^k)\| = 0\Big) = 1, \quad \lim_{k\to\infty}\mathbb{E}\big[\|\nabla F(\mathbf{x}_i^k)\|^2\big] = 0.$$

Moreover, if $\alpha = \overline{\alpha}_1$, ***GT-SAGA*** *achieves an* ϵ*-accurate stationary point in*

$$\mathcal{O}\left(\frac{EL(F(\overline{\mathbf{x}}^0) - F^*)}{\epsilon} + \frac{\lambda^2(1-\lambda^2)\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n\epsilon}\right) \tag{3.63}$$

iterations, where E is given by

$$E := \max\left\{\frac{m^{2/3}}{n^{1/3}}, 1, \frac{\lambda}{(1-\lambda)^2}, \frac{\lambda^{1/2}m^{1/2}}{(1-\lambda)^{3/4}}\right\}$$

and $\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2 = \sum_{i=1}^n \|\nabla f_i(\overline{\mathbf{x}}^0)\|^2$.

Theorem 3.3.1 is proved in Subsection 3.3.5.6. We discuss its implications in the following remarks.

Remark 3.3.3 (**Effect of the function heterogeneity**). We note that $\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2/n$ in the second term of (3.63) can be viewed as a measure of heterogeneity among the local functions. In particular, when all local functions are identical such that $f_i = f_r = F, \forall i, r \in \mathcal{V}$, this term diminishes, i.e., it can be shown that $\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2/n = \|\nabla F(\overline{\mathbf{x}}^0)\|^2 \leq 2L(F(\overline{\mathbf{x}}^0) - F^*)$. On the other hand, when the local functions are significantly different, $\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2/n$ can be fairly large compared with $L(F(\overline{\mathbf{x}}^0) - F^*)$. Based on Theorem 3.3.1, it is important to note that the effect of the function heterogeneity $\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2/n$ on the convergence rate of **GT-SAGA** is decoupled from E, the effect of the local batch size m and the network spectral gap $1-\lambda$. It is further interesting to observe that the heterogeneity effect diminishes when the network is sufficiently either well-connected or weakly-connected. In other words, the function heterogeneity effect is dominated by the network effect in these two extreme cases of interest.

We next view Theorem 3.3.1 in two different regimes.

Remark 3.3.4 (**Big-data regime**). We first consider a big-data regime that is often applicable in data centers, where the local batch size m is relatively large compared with the network spectral gap inverse $(1-\lambda)^{-1}$ and the number of the nodes n. In particular, if m large enough such that

$$\max\left\{1, \frac{\lambda}{(1-\lambda)^2}, \frac{\lambda^{1/2}m^{1/2}}{(1-\lambda)^{3/4}}\right\} \lesssim \frac{m^{2/3}}{n^{1/3}}, \tag{3.64}$$

Theorem 3.3.1 results into an iteration complexity of

$$\mathcal{O}\left(\frac{m^{2/3}L(F(\overline{\mathbf{x}}^0) - F^*)}{n^{1/3}\epsilon} + \frac{\lambda^2(1-\lambda^2)\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n\epsilon}\right). \tag{3.65}$$

We emphasize that the first term in (3.65) matches the iteration complexity of the centralized SAGA with a minibatch size n [60], as **GT-SAGA** computes n component gradients across the nodes in parallel at each iteration. We note that under the big-data condition (3.64), it typically holds that $\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2/n \lesssim m^{2/3}L(F(\overline{\mathbf{x}}^0) - F^*)/n^{1/3}$, i.e., the first term dominates the second term in (3.65). Therefore, **GT-SAGA** in this regime achieves a non-asymptotic linear speedup, i.e., the total number of component gradient computations required at each node to achieve an ϵ-accurate stationary point is reduced by a factor of $1/n$, compared with the centralized minibatch SAGA that operates on a single machine.

Remark 3.3.5 (**Large-scale network regime**). We now consider the case where a large number of nodes are weakly connected, a scenario that commonly appears in sensor networks, robotic swarms, and ad hoc IoT (Internet of Things) networks. In this case, the number of the nodes n and the network spectral gap inverse $(1-\lambda)^{-1}$ are relatively large in comparison with the local batch size m. In particular, if

$$\max\left\{1, \frac{m^{2/3}}{n^{1/3}}, \frac{\lambda^{1/2}m^{1/2}}{(1-\lambda)^{3/4}}\right\} \lesssim \frac{\lambda}{(1-\lambda)^2}, \tag{3.66}$$

then the component gradient computation complexity at each node of **GT-SAGA**, according to Theorem 3.3.1, becomes

$$\mathcal{O}\left(\frac{\lambda L(F(\overline{\mathbf{x}}^0)-F^*)}{(1-\lambda)^2\epsilon} + \frac{\lambda^2(1-\lambda^2)\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n\epsilon}\right). \tag{3.67}$$

We note that the component gradient complexity at each node of GT-SARAH [141], the state-of-the-art decentralized non-convex variance-reduced method, in this regime is

$$\mathcal{O}\left(\frac{\lambda}{(1-\lambda)^2\epsilon}\left(L(F(\overline{\mathbf{x}}^0)-F^*) + \frac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n}\right)\right). \tag{3.68}$$

Comparing (3.68) to (3.67), we observe that **GT-SARAH**, unlike **GT-SAGA**, does not achieve a separation between the dependence of the network spectral gap $1-\lambda$ and the function heterogeneity measure $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2/n$ on the convergence rate. We hence conclude that **GT-SAGA** outperforms **GT-SARAH** if the network is weakly connected and the local functions are largely heterogeneous, i.e., when $1-\lambda$ is small and $\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2/n$ is large. Moreover, we recall from Remark 3.3.1 that **GT-SAGA** is single-timescale and thus is much easier to implement than the two-timescale **GT-SARAH** over large-scale networks. We also emphasize that the storage requirement of **GT-SAGA** in this regime is significantly relaxed since the data samples are distributed across a large network, leading to a small local batch size m at each node.

3.3.3.2 Smooth non-convex functions under PL condition

Theorem 3.3.2. *Let Assumptions 3.3.1, 3.3.2, 3.3.3, and 3.3.4 hold. If the step-size α of **GT-SAGA** satisfies $0 < \alpha \leq \overline{\alpha}_2$, where*

$$\overline{\alpha}_2 := \min\left\{\frac{(1-\lambda^2)^2}{55\lambda L}, \frac{1-\lambda^2}{13\lambda\kappa^{1/4}L}, \frac{(1-\lambda^2)^3}{388\lambda^2 nL}, \frac{n^{1/3}}{10.5m^{2/3}\kappa^{1/3}L}, \frac{1}{36L}, \frac{1-\lambda^2}{2\mu}, \frac{1}{4m\mu}\right\},$$

then all nodes converge linearly at the rate $\mathcal{O}((1-\mu\alpha)^k)$ to a global minimizer of F. In particular, if $\alpha = \overline{\alpha}_2$, then all nodes agree on an ϵ-accurate global minimizer in

$$\mathcal{O}\left(\max\left\{Q_{opt}, Q_{net}\right\}\log\frac{1}{\epsilon}\right)$$

iterations, where Q_{opt} and Q_{net} are given respectively by

$$Q_{opt} := \max\left\{\frac{m^{2/3}\kappa^{4/3}}{n^{1/3}}, \kappa, m\right\},$$
$$Q_{net} := \max\left\{\frac{\lambda\kappa}{(1-\lambda)^2}, \frac{\lambda\kappa^{5/4}}{1-\lambda}, \frac{\lambda^2 n\kappa}{(1-\lambda)^3}, \frac{1}{1-\lambda}\right\}.$$

Theorem 3.3.2 is proved in Section 3.3.5.7. The following remarks are in place.

Remark 3.3.6 (**Linear rate under the global PL condition**). Theorem 3.3.2 shows that **GT-SAGA** linearly converges to an optimal solution when the global F additionally satisfies the PL condition. This is the first linear rate result for decentralized variance-reduced methods under the PL condition while the existing ones require strong convexity, e.g., [22–24, 120, 142]. A notable feature of the linear rate in Theorem 3.3.2 is that the effects of the local batch size m and the network spectral gap $1-\lambda$ are decoupled. Hence, in a big-data regime where the local batch size m is sufficiently large such that $Q_{\rm net} \lesssim Q_{\rm opt}$, **GT-SAGA** achieves a network topology-independent rate of $\mathcal{O}(Q_{\rm opt}\log\frac{1}{\epsilon})$. In addition, we note that Theorem 3.3.2 implies the linear rate of **GT-SAGA** in the almost sure sense under the PL condition, by Chebyshev's inequality and the Borel-Cantelli lemma; see Lemma 7 in [24] for details.

Remark 3.3.7 (**Comparison with other decentralized gradient methods**). When the local batch size m is relatively large, the linear rate of **GT-SAGA** improves that of the existing decentralized batch gradient methods [4, 143, 144] under the PL condition in terms of the component gradient computation complexity. Moreover, decentralized online stochastic gradient methods, e.g., [4,148], only exhibit sublinear rate under the PL condition due to the persistent variances of the stochastic gradients. Therefore, **GT-SAGA** achieves faster convergence under the PL condition compared with the existing decentralized methods, demonstrating the advantage of the employed SAGA variance reduction scheme that is able to exploit the finite-sum structure of local functions.

Remark 3.3.8 (**Improved convergence results for the centralized minibatch SAGA**). When $\lambda = 0$, i.e., when the underlying network is a complete graph whose weight matrix can be easily chosen as $\underline{\mathbf{W}} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, **GT-SAGA** reduces to the centralized minibatch SAGA and achieves the linear rate of $\mathcal{O}(Q_{\rm opt}\log\frac{1}{\epsilon})$. Hence, a special case of Theorem 3.3.2, i.e, $\lambda = 0$, provides the first linear rate result under the PL condition for the centralized SAGA. Indeed, the existing linear rate results [60, 145] under the PL condition are only applicable to a modified SAGA that *periodically restarts* $\mathcal{O}(\log\frac{1}{\epsilon})$ times with the output of each cycle being selected randomly from the past iterates in this cycle. This procedure is not feasible particularly in decentralized settings. In contrast, the linear rate shown Theorem 3.3.2 is on the *last iterate* of the original SAGA without periodic restarting and sampling.

Table 3.4: Datasets used in numerical experiments, available at https://www.openml.org/.

Dataset	**train** ($N = nm$)	**dimension** (p)
nomao	30,000	119
a9a	48,800	124
w8a	60,000	300
KDD98	80,000	478
covertype	100,000	55
MiniBooNE	100,000	51
BNG(sonar)	100,000	61

3.3.4 Numerical experiments

In this section, we present numerical simulations to illustrate our main theoretical results. The network topologies of interest are undirected ring, undirected 2D-grid, directed exponential, undirected geometric, and complete graphs; see [24, 27, 41] for details of these graphs. The doubly stochastic weights are set to be equal for the ring and exponential graphs, and are generated by the lazy Metropolis rule for the grid and geometric graphs. We manually optimize the parameters of all algorithms in all experiments for their best performance.

3.3.4.1 Non-convex binary classification

In this subsection, we consider a decentralized non-convex generalized linear model for binary classification. In view of Problem (3.1), each component cost $f_{i,j}$ is defined as [146]

$$f_{i,j}(\mathbf{x}) := \ell\big(\xi_{i,j}\mathbf{x}^\top \boldsymbol{\theta}_{i,j}\big), \quad \ell(u) := \left(1 - \frac{1}{1+\exp(-u)}\right)^2,$$

where $\boldsymbol{\theta}_{i,j} \in \mathbb{R}^p$ is the j-th data vector at the i-th node, $\xi_{i,j} \in \{-1, +1\}$ is the label of $\boldsymbol{\theta}_{i,j}$, and $\ell : \mathbb{R} \to \mathbb{R}$ is a $\frac{4}{3}$-smooth non-convex loss. We normalize each data to be $\|\boldsymbol{\theta}_{i,j}\| = 1, \forall i, j$. Since $\nabla^2 f_{i,j}(\mathbf{x}) = \ell''(\xi_{i,j}\mathbf{x}^\top \boldsymbol{\theta}_{i,j})\boldsymbol{\theta}_{i,j}\boldsymbol{\theta}_{i,j}^\top$, it can be verified that $\|\nabla^2 f_{i,j}(\mathbf{x})\| = |\ell''(\xi_{i,j}\mathbf{x}^\top \boldsymbol{\theta}_{i,j})| \leq \frac{4}{3}$. Hence each component cost $f_{i,j}$ is non-convex and $\frac{4}{3}$-smooth. We measure the performance of the algorithms in question in terms of the decrease of the stationary gap $\|\nabla F(\overline{\mathbf{x}})\|$ versus epochs, where $\overline{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ for $\mathbf{x}_i$ being the model at node i and each epoch represents m component gradient evaluations at each node. All nodes start from a vector randomly generated from the standard Gaussian distribution. The statistics of the datasets used in the experiments are provided in Table 3.4.

- **Big data regime.** We first test the convergence behavior of `GT-SAGA` in the big data regime by uniformly distributing the KDD98, covertype, MiniBooNE, and BNG(sonar) datasets over a network

of $n = 20$ nodes. We consider four different network topologies with decreasing sparsity, i.e., the undirected ring, undirected 2D-grid, directed exponential, and complete graph; their corresponding second largest singular values of the weight matrices are $\lambda = 0.98, 0.97, 0.6, 0$, respectively. It can be verified that the big data condition (3.64) holds. The experimental results are shown in Fig. 3.4, where we observe that the convergence rate of **GT-SAGA** is independent of the network topology in this big data regime; see Remark 3.3.4.

- **Large-scale network regime.** We next compare the performance of **GT-SAGA** with DSGD [2] and GT-SARAH [141] in the large-scale network regime. To this aim, we generate a sparse geometric graph of $n = 200$ nodes with $\lambda \approx 0.99$ and uniformly distribute the nomao, a9a, w8a, and BNG(sonar) datasets over the nodes. It can be verified that the large-scale network condition (3.66) holds. The numerical results are presented in Fig. 3.5: the first three plots show that **GT-SAGA** achieves the best performance among the algorithms in comparison, while the last plot shows that the convergence rate of **GT-SAGA** is dependent on the network topology in this large-scale network regime; see Remark 3.3.5.

- **Robustness to heterogeneous data.** We now make the data distributions across the nodes significantly heterogeneous by letting each node only have data samples of one label, so that no node can train a valid classification model only from its local data. We compare the performance of **GT-SAGA** under heterogeneous and homogeneous distribution of the nomao dataset. We consider a well-connected graph, i.e., the 20-node exponential graph, and a weakly-connected graph, i.e., the 200-node geometric graph. The numerical results are shown in Fig. 3.6, where we observe that the convergence rate of **GT-SAGA** is not affected by the data heterogeneity over both graphs; see Remark 3.3.3.

3.3.4.2 Synthetic functions that satisfy the PL condition

Finally, we verify the linear rate of **GT-SAGA** when the global function F satisfies the PL condition. Specifically, we choose each component function $f_{i,j} : \mathbb{R} \to \mathbb{R}$ as

$$f_{i,j}(x) = x^2 + 3\sin^2(x) + a_{i,j}\cos(x) + b_{i,j}x,$$

where $\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j} = 0$ and $\sum_{i=1}^{n}\sum_{j=1}^{m} b_{i,j} = 0$ such that $a_{i,j} \neq 0$, $b_{i,j} \neq 0$, $\forall i, j$. This formulation hence leads to the global function $F(x) = x^2 + 3\sin^2(x)$. It can be verified that F is non-convex and satisfies the PL condition [147]. Note that each $f_{i,j}$ is nonlinear and highly deviated from F; see the last three plots in Fig. 3.7 for a comparison of local and global geometries. We use the 20-node exponential graph and set $m = 5$. It can be observed from the first plot in Fig. 3.7 that **GT-SAGA** achieves linear rate to the optimal solution, while DSGD converges to an inexact solution; see Remark 3.3.7.

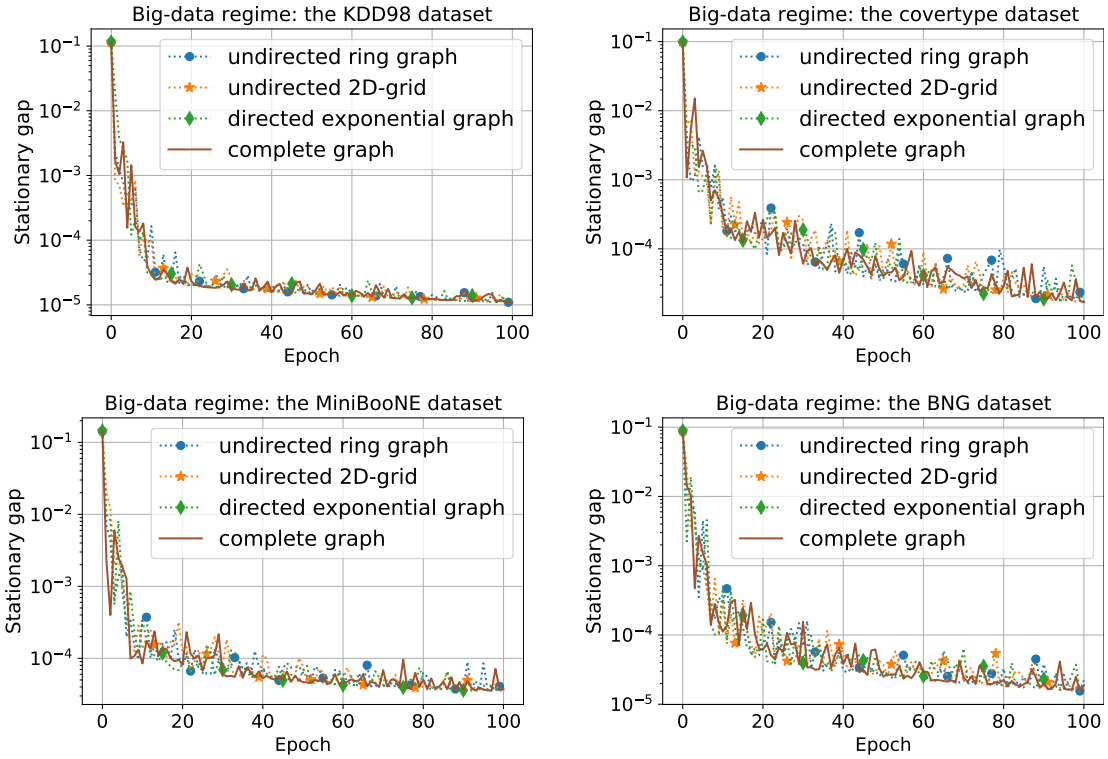

Figure 3.4: Big data regime: the network topology-independent convergence rate of **GT-SAGA** on the KDD98, covertype, MiniBooNE, and BNG(sonar) datasets.

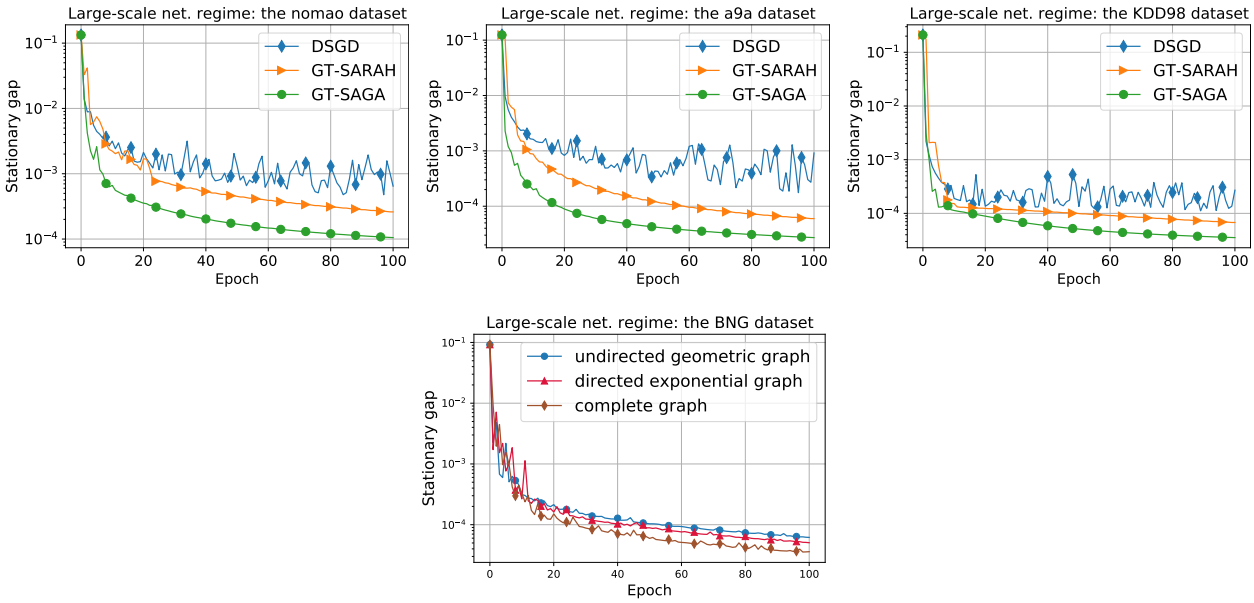

Figure 3.5: Large-scale network regime: (i) the first three plots present the performance comparison between **GT-SAGA**, DSGD, and GT-SARAH on the nomao, a9a, and KDD98 datasets; (ii) the last plot presents the performance of **GT-SAGA** over different graph topologies in this regime on the BNG(sonar) dataset.

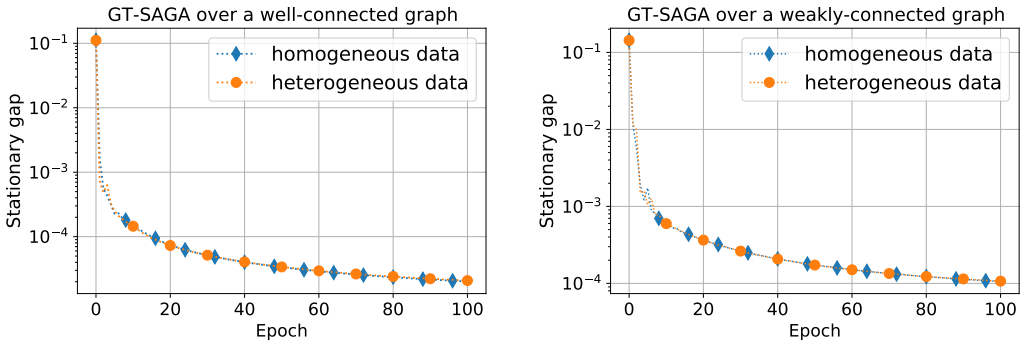

Figure 3.6: Robustness of **GT-SAGA** to heterogeneous data over well- and weakly-connected graphs on the nomao dataset.

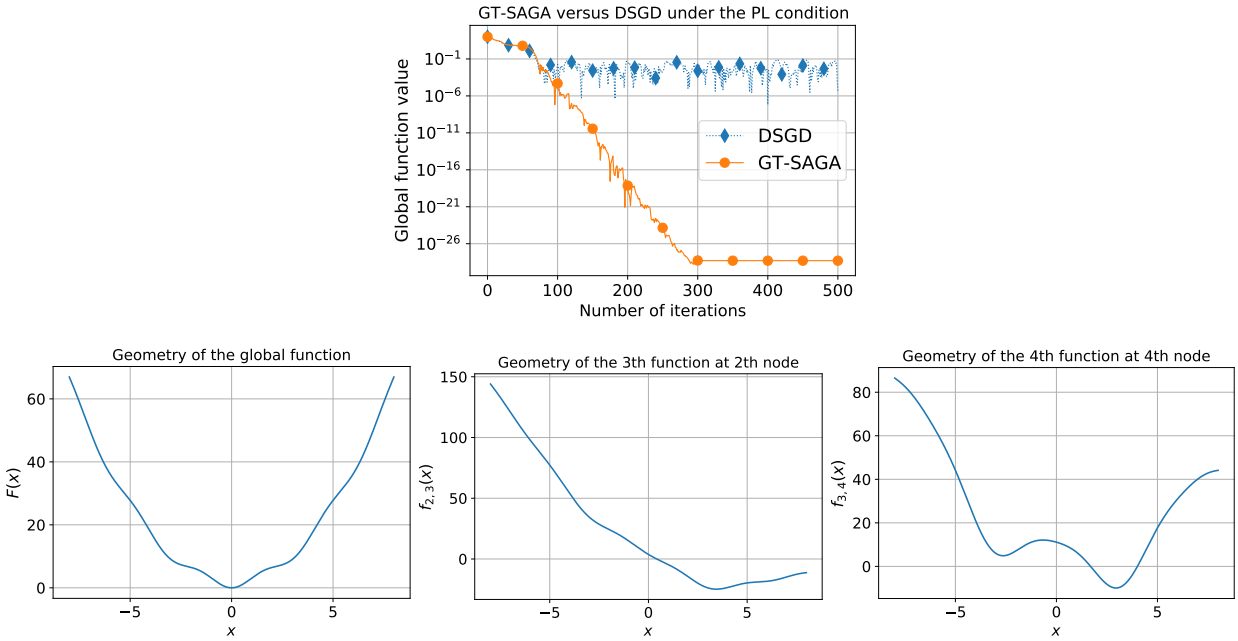

Figure 3.7: The PL condition: (i) the first plot presents the performance comparison between **GT-SAGA** and DSGD when the global function satisfies the PL condition; (ii) the last three plots present the geometry comparison of the global and local component functions.

3.3.5 Convergence analysis

In this section, we present the convergence analysis of **GT-SAGA**, i.e., the sublinear convergence for general smooth non-convex functions and the linear convergence when the global function F additionally satisfies the PL condition. Throughout this section, we assume Assumption 3.3.1, 3.3.2, and 3.3.4 hold without explicitly stating them; we only assume Assumption 3.3.3 hold in Subsection 3.3.5.7. In Subsections 3.3.5.2-3.3.5.5, we establish key relationships between several important quantities, based on which the proofs of Theorem 3.3.1 and 3.3.2 are derived in Subsections 3.3.5.6 and 3.3.5.7 respectively.

3.3.5.1 Preliminaries

GT-SAGA can be written in the following form: $\forall k \geq 0$,

$$\mathbf{y}^{k+1} = \mathbf{W}\left(\mathbf{y}^k + \mathbf{g}^k - \mathbf{g}^{k-1}\right), \tag{3.69a}$$

$$\mathbf{x}^{k+1} = \mathbf{W}\left(\mathbf{x}^k - \alpha \mathbf{y}^{k+1}\right), \tag{3.69b}$$

where $\mathbf{x}^k, \mathbf{y}^k, \mathbf{g}^k$ are random vectors in $\mathbb{R}^{np}$ that concatenate all local states $\{\mathbf{x}_i^k\}_{i=1}^n$, gradient trackers $\{\mathbf{y}_i^k\}_{i=1}^n$, local SAGA estimators $\{\mathbf{g}_i^k\}_{i=1}^n$, respectively, and $\mathbf{W} = \underline{\mathbf{W}} \otimes \mathbf{I}_p$. We denote $\mathcal{F}^k$ as the filtration of **GT-SAGA**, i.e., $\forall k \geq 1$,

$$\mathcal{F}^k := \sigma\left(\left\{\tau_i^t, s_i^t : i \in \mathcal{V}, t \leq k-1\right\}\right),$$

and $\mathcal{F}^0$ is the trivial σ-algebra. It can be verified that $\mathbf{x}^k$, $\mathbf{y}^k$ and $\mathbf{z}_{i,j}^k, \forall i, j$, are $\mathcal{F}^k$-measurable and $\mathbf{g}^k$ is $\mathcal{F}^{k+1}$-measurable for all $k \geq 0$. We use $\mathbb{E}[\cdot|\mathcal{F}^k]$ to denote the conditional expectation with respect to $\mathcal{F}^k$. For the ease of exposition, we introduce the following quantities:

$$\begin{aligned}
&\mathbf{J} := \left(\mathbf{1}_n \mathbf{1}_n^\top / n\right) \otimes \mathbf{I}_p, \\
&\nabla \mathbf{f}(\mathbf{x}^k) = \left[\nabla f_1(\mathbf{x}_1^k)^\top, \cdots, \nabla f_n(\mathbf{x}_n^k)^\top\right]^\top, \\
&\overline{\nabla \mathbf{f}}(\mathbf{x}^k) = \left(\mathbf{1}_n^\top \otimes \mathbf{I}_p / n\right) \nabla \mathbf{f}(\mathbf{x}^k), \qquad \overline{\mathbf{x}}^k = \left(\mathbf{1}_n^\top \otimes \mathbf{I}_p / n\right) \mathbf{x}^k, \\
&\overline{\mathbf{y}}^k = \left(\mathbf{1}_n^\top \otimes \mathbf{I}_p / n\right) \mathbf{y}^k, \qquad \overline{\mathbf{g}}^k = \left(\mathbf{1}_n^\top \otimes \mathbf{I}_p / n\right) \mathbf{g}^k.
\end{aligned}$$

We assume $\overline{\mathbf{x}}^0 \in \mathbb{R}^p$ is constant and hence all random variables generated by **GT-SAGA** have bounded second moment. The following lemma lists several well-known facts in the context of gradient tracking and SAGA estimators, which may be found in [5, 24, 55–57].

Lemma 3.3.1. *The following relationships hold.*

(a) $\forall \mathbf{x} \in \mathbb{R}^{np}$, $\|\mathbf{W}\mathbf{x} - \mathbf{J}\mathbf{x}\| \leq \lambda \|\mathbf{x} - \mathbf{J}\mathbf{x}\|$.

(b) $\overline{\mathbf{y}}^{k+1} = \overline{\mathbf{g}}^k, \forall k \geq 0$.

(c) $\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k) - \nabla F(\overline{\mathbf{x}}^k)\|^2 \leq \frac{L^2}{n} \|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2, \forall k \geq 0$.

(d) $\mathbb{E}[\mathbf{g}_i^k | \mathcal{F}^k] = \nabla f_i(\mathbf{x}_i^k), \forall i \in \mathcal{V}, \forall k \geq 0$.

(e) $\|\nabla F(\mathbf{x})\|^2 \leq 2L\left(F(\mathbf{x}) - F^*\right)$.

Note that Lemma 3.3.1(e) is a consequence of the L-smoothness of the global function F and is only used in Subsection 3.3.5.7 while other statements in Lemma 3.3.1 are frequently utilized throughout the analysis. The next lemma states some standard inequalities on the network consensus error [4, 24].

Lemma 3.3.2. *The following inequality holds:* $k \geq 0$,

$$\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\|^2 \leq \tfrac{1+\lambda^2}{2}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \tfrac{2\alpha^2\lambda^2}{1-\lambda^2}\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2. \tag{3.70}$$

$$\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\|^2 \leq 2\lambda^2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 2\alpha^2\lambda^2\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2. \tag{3.71}$$

$$\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\| \leq \lambda\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\| + \alpha\lambda\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|. \tag{3.72}$$

3.3.5.2 Bounds on the variance of local SAGA estimators

In this subsection, we bound the variance of the local SAGA gradient estimators $\mathbf{g}_i^k$'s. For analysis purposes, we construct two auxiliary $\mathcal{F}^k$-adapted sequences: $\forall i \in \mathcal{V}$, $\forall k \geq 0$,

$$t_i^k := \frac{1}{m}\sum_{j=1}^m \|\overline{\mathbf{x}}^k - \mathbf{z}_{i,j}^k\|^2, \qquad t^k := \frac{1}{n}\sum_{i=1}^n t_i^k.$$

These two sequences are essential in the convergence analysis. We note that t^k measures the average distance between the mean state $\overline{\mathbf{x}}^k$ of the networked nodes and the latest iterates $\mathbf{z}_{i,j}^k$'s where the component gradients were computed at iteration k in the gradient tables. Intuitively, t^k goes to 0 as all nodes in **GT-SAGA** reach consensus on a stationary point. We will establish a contraction argument in t^k in Subsection 3.3.5.4. In the following lemma, we show that the variance of $\mathbf{g}_i^k$ may be bounded by the network consensus error and t^k.

Lemma 3.3.3. *The following inequality holds:* $\forall k \geq 0$,

$$\mathbb{E}\big[\|\mathbf{g}^k - \nabla\mathbf{f}(\mathbf{x}^k)\|^2|\mathcal{F}^k\big] \leq 2L^2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 2nL^2t^k, \tag{3.73}$$

$$\mathbb{E}\big[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\big] \leq \frac{2L^2}{n^2}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \frac{2L^2}{n}t^k + \|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2. \tag{3.74}$$

Proof. We denote $\widehat{\nabla}_i^k := \nabla f_{i,\tau_i^k}(\mathbf{x}_i^k) - \nabla f_{i,\tau_i^k}(\mathbf{z}_{i,\tau_i^k}^k)$ for ease of exposition. Observe from Algorithm 4 that

$$\mathbb{E}\big[\widehat{\nabla}_i^k|\mathcal{F}^k\big] = \nabla f_i(\mathbf{x}_i^k) - \frac{1}{m}\sum_{j=1}^m \nabla f_{i,j}(\mathbf{z}_{i,j}^k) \tag{3.75}$$

for all k and i. In light of (3.75), we bound the variance of $\mathbf{g}_i^k$ in the following: $\forall k \geq 0, \forall i \in \mathcal{V}$,

$$\begin{aligned}
\mathbb{E}\big[\|\mathbf{g}_i^k - \nabla f_i(\mathbf{x}_i^k)\|^2|\mathcal{F}^k\big] &= \mathbb{E}\big[\|\widehat{\nabla}_i^k - \mathbb{E}\big[\widehat{\nabla}_i^k|\mathcal{F}^k\big]\|^2|\mathcal{F}^k\big] \\
&\overset{(i)}{\leq} \mathbb{E}\big[\|\widehat{\nabla}_i^k\|^2\big|\mathcal{F}^k\big] \\
&= \mathbb{E}\bigg[\sum_{j=1}^m \mathbb{1}_{\{\tau_i^k=j\}}\big\|\nabla f_{i,j}(\mathbf{x}_i^k) - \nabla f_{i,j}(\mathbf{z}_{i,j}^k)\big\|^2\big|\mathcal{F}^k\bigg] \\
&\overset{(ii)}{=} \frac{1}{m}\sum_{j=1}^m \big\|\nabla f_{i,j}(\mathbf{x}_i^k) - \nabla f_{i,j}(\mathbf{z}_{i,j}^k)\big\|^2 \\
&\overset{(iii)}{\leq} \frac{L^2}{m}\sum_{j=1}^m \big\|\mathbf{x}_i^k - \mathbf{z}_{i,j}^k\big\|^2 \\
&\leq 2L^2\big\|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\big\|^2 + 2L^2t_i^k.
\end{aligned} \tag{3.76}$$

where (i) the conditional variance decomposition, (ii) uses that $\|\nabla f_{i,j}(\mathbf{x}_i^k) - \nabla f_{i,j}(\mathbf{z}_{i,j}^k)\|^2$ is $\mathcal{F}^k$-measurable and that τ_i^k is independent of $\mathcal{F}^k$, and (iii) uses the L-smoothness of each $f_{i,j}$. Summing up (3.76) over i from 1 to n gives (3.73). Towards (3.74), we have: $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\big[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\big] &\overset{(i)}{=} \mathbb{E}\big[\|\overline{\mathbf{g}}^k - \overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2|\mathcal{F}^k\big] + \|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 \\ &\overset{(ii)}{=} \frac{1}{n^2}\mathbb{E}\big[\|\mathbf{g}^k - \nabla \mathbf{f}(\mathbf{x}^k)\|^2|\mathcal{F}^k\big] + \|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2, \end{aligned} \tag{3.77}$$

where (i) uses that $\mathbb{E}[\overline{\mathbf{g}}^k|\mathcal{F}^k] = \overline{\nabla \mathbf{f}}(\mathbf{x}^k)$ and that $\overline{\nabla \mathbf{f}}(\mathbf{x}^k)$ is $\mathcal{F}^k$-measurable while (ii) uses that, whenever $i \neq j$, $\mathbb{E}[\langle \mathbf{g}_i^k - \nabla f_i(\mathbf{x}_i^k), \mathbf{g}_j^k - \nabla f_j(\mathbf{x}_j^k)\rangle|\mathcal{F}^k] = 0$, since τ_i^k is independent of $\sigma(\sigma(\tau_j^k), \mathcal{F}^k)$ and $\mathbb{E}[\mathbf{g}^k|\mathcal{F}^k] = \nabla \mathbf{f}(\mathbf{x}^k)$. The proof follows by applying (3.73) to (3.77). □

3.3.5.3 A descent inequality

In this subsection, we provide a key descent inequality that characterizes the expected decrease of the global function value at each iteration of **GT-SAGA**.

Lemma 3.3.4. *If* $0 < \alpha \leq \frac{1}{2L}$, *then* $\forall k \geq 0$,

$$\mathbb{E}\big[F(\overline{\mathbf{x}}^{k+1})|\mathcal{F}^k\big] \leq F(\overline{\mathbf{x}}^k) - \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}^k)\|^2 - \frac{\alpha}{4}\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 + \frac{\alpha L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \frac{\alpha^2 L^3}{n}t^k.$$

Proof. Since F is L-smooth, we have [10]: $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{3.78}$$

We multiply (3.69b) by $\frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)$ and use Lemma 3.3.1(b) to obtain:

$$\overline{\mathbf{x}}^{k+1} = \overline{\mathbf{x}}^k - \alpha\overline{\mathbf{y}}^{k+1} = \overline{\mathbf{x}}^k - \alpha\overline{\mathbf{g}}^k, \qquad \forall k \geq 0.$$

Setting $\mathbf{y} = \overline{\mathbf{x}}^{k+1}$ and $\mathbf{x} = \overline{\mathbf{x}}^k$ in (3.78) obtains: $\forall k \geq 0$,

$$F(\overline{\mathbf{x}}^{k+1}) \leq F(\overline{\mathbf{x}}^k) - \alpha\langle \nabla F(\overline{\mathbf{x}}^k), \overline{\mathbf{g}}^k\rangle + \frac{\alpha^2 L}{2}\|\overline{\mathbf{g}}^k\|^2. \tag{3.79}$$

Conditioning (3.79) with respect to $\mathcal{F}^k$, since $\nabla F(\overline{\mathbf{x}}^k)$ is $\mathcal{F}^k$-measurable, we have:

$$\mathbb{E}\left[F(\overline{\mathbf{x}}^{k+1})|\mathcal{F}^k\right] \leq F(\overline{\mathbf{x}}^k) - \alpha\left\langle \nabla F(\overline{\mathbf{x}}^k), \overline{\nabla \mathbf{f}}(\mathbf{x}^k)\right\rangle + \frac{\alpha^2 L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right]. \tag{3.80}$$

Using $2\langle \mathbf{a}, \mathbf{b}\rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2$,$\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, in (3.80), we obtain: $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\big[F(\overline{\mathbf{x}}^{k+1})|\mathcal{F}^k\big] \leq F(\overline{\mathbf{x}}^k) &- \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}^k)\|^2 - \frac{\alpha}{2}\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 \\ &+ \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}^k) - \overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 + \frac{\alpha^2 L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\right]. \end{aligned} \tag{3.81}$$

Applying Lemma 3.3.1(c) and (3.74) to (3.81), we have: $\forall k \geq 0$,

$$\mathbb{E}[F(\overline{\mathbf{x}}^{k+1})|\mathcal{F}^k] \leq F(\overline{\mathbf{x}}^k) - \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}^k)\|^2 - \frac{\alpha(1-\alpha L)}{2}\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 + \Big(\frac{\alpha L^2}{2n} + \frac{\alpha^2 L^3}{n^2}\Big)\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \frac{\alpha^2 L^3}{n} t^k. \tag{3.82}$$

The proof follows by the fact that if $0 < \alpha \leq \frac{1}{2L}$, we have $-\frac{\alpha(1-\alpha L)}{2} \leq -\frac{\alpha}{4}$ and $\frac{\alpha L^2}{2n} + \frac{\alpha^2 L^3}{n^2} \leq \frac{\alpha L^2}{n}$. □

Compared with the corresponding descent inequality for centralized batch gradient descent [10], Lemma 3.3.4 exhibits two additional bias terms, i.e., $\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|$ and t^k, that are due to the decentralized nature of the problem and sampling. To establish the convergence of **GT-SAGA**, we therefore bound these bias terms by $\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|$ and show that they are dominated by the descent effect $-\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|$.

3.3.5.4 Bounds on the auxiliary sequence t^k

In this subsection, we analyze the evolution of the auxiliary sequence t^k and establish useful bounds.

Lemma 3.3.5. *The following inequality holds: $\forall k \geq 0$,*

$$\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right] \leq \theta t^k + \Big(2\alpha^2 + \frac{\alpha}{\beta}\Big)\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 + \Big(\frac{2\alpha^2 L^2}{n} + \frac{2}{m}\Big)\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2,$$

where the parameter $\theta \in \mathbb{R}$ is given by

$$\theta := 1 - \frac{1}{m} + \alpha\beta + \frac{2\alpha^2 L^2}{n}, \tag{3.83}$$

and $\beta > 0$ is an arbitrary positive constant.

Proof. We define $\mathcal{A}^k := \sigma\left(\cup_{i=1}^n \sigma(\tau_i^k), \mathcal{F}^k\right)$ and clearly $\mathcal{F}^k \subseteq \mathcal{A}^k$. By the tower property of the conditional expectation, we have: $\forall i \in \mathcal{V}, \forall k \geq 0$,

$$\mathbb{E}\big[t_i^{k+1}|\mathcal{F}^k\big] = \frac{1}{m}\sum_{j=1}^m \mathbb{E}\big[\mathbb{E}\big[\|\overline{\mathbf{x}}^{k+1} - \mathbf{z}_{i,j}^{k+1}\|^2|\mathcal{A}^k\big]|\mathcal{F}^k\big]. \tag{3.84}$$

Since s_i^k is independent of $\mathcal{A}^k$ under Assumption 3.3.1, we have: $\forall i \in \mathcal{V}, \forall j \in \{1, \cdots, m\}, k \geq 0$,

$$\mathbb{E}\Big[\mathbb{1}_{\{s_i^k = j\}}|\mathcal{A}^k\Big] = \frac{1}{m} \qquad \text{and} \qquad \mathbb{E}\Big[\mathbb{1}_{\{s_i^k \neq j\}}|\mathcal{A}^k\Big] = 1 - \frac{1}{m}. \tag{3.85}$$

In light of (3.85), we have: $\forall i \in \mathcal{V}, \forall j \in \{1, \cdots, m\}, k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{x}}^{k+1} - \mathbf{z}_{i,j}^{k+1}\|^2|\mathcal{A}^k\right] \\
&= \mathbb{E}\Big[\Big\|\overline{\mathbf{x}}^{k+1} - \Big(\mathbb{1}_{\{s_i^k = j\}}\mathbf{x}_i^k + \mathbb{1}_{\{s_i^k \neq j\}}\mathbf{z}_{i,j}^k\Big)\Big\|^2\Big|\mathcal{A}^k\Big] \\
&= \mathbb{E}\big[\|\overline{\mathbf{x}}^{k+1}\|^2|\mathcal{A}^k\big] + \mathbb{E}\Big[\Big\|\mathbb{1}_{\{s_i^k = j\}}\mathbf{x}_i^k + \mathbb{1}_{\{s_i^k \neq j\}}\mathbf{z}_{i,j}^k\Big\|^2\Big|\mathcal{A}^k\Big] - 2\mathbb{E}\Big[\Big\langle \overline{\mathbf{x}}^{k+1}, \mathbb{1}_{\{s_i^k = j\}}\mathbf{x}_i^k + \mathbb{1}_{\{s_i^k \neq j\}}\mathbf{z}_{i,j}^k\Big\rangle\Big|\mathcal{A}^k\Big] \\
&\overset{(i)}{=} \|\overline{\mathbf{x}}^{k+1}\|^2 - 2\Big\langle \overline{\mathbf{x}}^{k+1}, \frac{1}{m}\mathbf{x}_i^k + \Big(1 - \frac{1}{m}\Big)\mathbf{z}_{i,j}^k\Big\rangle + \frac{1}{m}\|\mathbf{x}_i^k\|^2 + \Big(1 - \frac{1}{m}\Big)\|\mathbf{z}_{i,j}^k\|^2, \\
&= \frac{1}{m}\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}_i^k\|^2 + \Big(1 - \frac{1}{m}\Big)\|\overline{\mathbf{x}}^{k+1} - \mathbf{z}_{i,j}^k\|^2
\end{aligned} \tag{3.86}$$

where (i) uses (3.85) and that $\overline{\mathbf{x}}^{k+1}, \mathbf{x}_i^k$, and $\mathbf{z}_{i,j}^k$ are $\mathcal{A}^k$-measurable. Using (3.86) in (3.84), we obtain the following: $\forall i \in \mathcal{V}, \forall k \geq 0$,

$$\mathbb{E}\big[t_i^{k+1}|\mathcal{F}^k\big] = \frac{1}{m}\mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}_i^k\right\|^2 \Big| \mathcal{F}^k\right] + \Big(1 - \frac{1}{m}\Big)\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left[\left\|\overline{\mathbf{x}}^{k+1} - \mathbf{z}_{i,j}^k\right\|^2 \Big| \mathcal{F}^k\right]. \tag{3.87}$$

We next bound the two terms on the RHS of (3.87) separately. For the first term, we have: $\forall i \in \mathcal{V}$, $k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\big[\|\overline{\mathbf{x}}^{k+1} - \mathbf{x}_i^k\|^2|\mathcal{F}^k\big] \\
&= \mathbb{E}\big[\|\overline{\mathbf{x}}^{k+1} - \overline{\mathbf{x}}^k + \overline{\mathbf{x}}^k - \mathbf{x}_i^k\|^2|\mathcal{F}^k\big] \\
&= \alpha^2\mathbb{E}\big[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\big] - 2\big\langle\alpha\overline{\nabla\mathbf{f}}(\mathbf{x}^k), \overline{\mathbf{x}}^k - \mathbf{x}_i^k\big\rangle + \|\overline{\mathbf{x}}^k - \mathbf{x}_i^k\|^2 \\
&\leq \alpha^2\mathbb{E}\big[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\big] + \alpha^2\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 + 2\|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2,
\end{aligned} \tag{3.88}$$

where the last line uses the Cauchy-Schwarz inequality. Towards the second term on the RHS of (3.87), we have: $\forall i \in \mathcal{V}$, $j \in \{1, \cdots, m\}$, $\forall k \geq 0$, $\forall \beta > 0$,

$$\begin{aligned}
&\mathbb{E}\big[\|\overline{\mathbf{x}}^{k+1} - \mathbf{z}_{i,j}^k\|^2|\mathcal{F}^k\big] \\
&= \mathbb{E}\big[\|\overline{\mathbf{x}}^{k+1} - \overline{\mathbf{x}}^k + \overline{\mathbf{x}}^k - \mathbf{z}_{i,j}^k\|^2\big|\mathcal{F}^k\big] \\
&= \alpha^2\mathbb{E}\big[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\big] - 2\alpha\big\langle\overline{\nabla\mathbf{f}}(\mathbf{x}^k), \overline{\mathbf{x}}^k - \mathbf{z}_{i,j}^k\big\rangle + \|\overline{\mathbf{x}}^k - \mathbf{z}_{i,j}^k\|^2 \\
&\leq \alpha^2\mathbb{E}\big[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\big] + (1+\alpha\beta)\|\overline{\mathbf{x}}^k - \mathbf{z}_{i,j}^k\|^2 + \frac{\alpha}{\beta}\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2,
\end{aligned} \tag{3.89}$$

where the last line uses Young's inequality. We apply (3.88) and (3.89) to (3.87) to obtain: $\forall i \in \mathcal{V}$, $\forall k \geq 0$,

$$\begin{aligned}
\mathbb{E}\big[t_i^{k+1}|\mathcal{F}^k\big] \leq{}& \Big(1 - \frac{1}{m}\Big)(1+\alpha\beta)t_i^k + \alpha^2\mathbb{E}\big[\|\overline{\mathbf{g}}^k\|^2|\mathcal{F}^k\big] \\
&+ \frac{2}{m}\|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2 + \Big(\frac{\alpha^2}{m} + \Big(1 - \frac{1}{m}\Big)\frac{\alpha}{\beta}\Big)\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2.
\end{aligned} \tag{3.90}$$

We average (3.90) over i from 1 to n and use (3.74) in the resulting inequality to obtain: $\forall k \geq 0$,

$$\begin{aligned}
\mathbb{E}\big[t^{k+1}|\mathcal{F}^k\big] \leq{}& \Big(\frac{2\alpha^2L^2}{n} + \frac{2}{m}\Big)\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \left(\alpha^2 + \frac{\alpha^2}{m} + \Big(1 - \frac{1}{m}\Big)\frac{\alpha}{\beta}\right)\left\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\right\|^2 \\
&+ \left(\frac{2\alpha^2L^2}{n} + \Big(1 - \frac{1}{m}\Big)(1+\alpha\beta)\right)t^k.
\end{aligned} \tag{3.91}$$

We conclude by using $\frac{1}{m} + 1 \leq 2$ and $1 - \frac{1}{m} \leq 1$ in (3.91). □

Next, we specify some particular choices of β and the range of α in Lemma 3.3.5 to obtain useful bounds on the auxiliary sequence t^k. The following corollary shows that t^k has an intrinsic contraction property.

Corollary 3.3.1. *If* $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8mL}}$, *then* $\forall k \geq 0$,

$$\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right] \leq \Big(1 - \frac{1}{4m}\Big)t^k + 4m\alpha^2\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 + \frac{9}{4mn}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2.$$

Proof. We choose $\beta = \frac{1}{2m\alpha}$ in Lemma 3.3.5 to obtain: if $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, i.e., $\frac{2\alpha^2L^2}{n} \leq \frac{1}{4m}$, then

$$\theta = 1 - \frac{1}{m} + \alpha\beta + \frac{2\alpha^2L^2}{n} \leq 1 - \frac{1}{4m}. \tag{3.92}$$

$$2\alpha^2 + \frac{\alpha}{\beta} = 2\alpha^2 + 2m\alpha^2 \leq 4m\alpha^2. \tag{3.93}$$

$$\frac{2\alpha^2L^2}{n} + \frac{2}{m} \leq \frac{1}{4m} + \frac{2}{m} = \frac{9}{4m}. \tag{3.94}$$

We conclude by applying (3.92), (3.93), (3.94) to Lemma 3.3.5. □

The following corollary of Lemma 3.3.5 will be only used to bound $\mathbb{E}[\|\mathbf{g}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^k]$.

Corollary 3.3.2. *If* $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, *then* $\forall k \geq 0$,

$$\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right] \leq 2t^k + 3\alpha^2\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 + \frac{9}{4mn}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2.$$

Proof. Setting $\beta = 1/\alpha$ in Lemma 3.3.5, we have: if $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, i.e., $\frac{2\alpha^2L^2}{n} \leq \frac{1}{4m}$, then

$$\theta = 1 - \frac{1}{m} + \alpha\beta + \frac{2L^2\alpha^2}{n} \leq 2 \tag{3.95}$$

$$2\alpha^2 + \frac{\alpha}{\beta} = 3\alpha^2 \tag{3.96}$$

$$\frac{2\alpha^2L^2}{n} + \frac{2}{m} \leq \frac{1}{4m} + \frac{2}{m} = \frac{9}{4m}, \tag{3.97}$$

We conclude by applying (3.95), (3.96), (3.97) to Lemma 3.3.5. □

With the help of (3.71), (3.73) and Corollary 3.3.2, we provide an upper bound on $\mathbb{E}[\|\mathbf{g}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^k]$.

Lemma 3.3.6. *If* $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, *then* $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^k] \leq\ & 8.5L^2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 4nL^2t^k \\ & + 6n\alpha^2L^2\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 + 4\alpha^2L^2\mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2|\mathcal{F}^k\right].\end{aligned}$$

Proof. By the tower property of the conditional expectation, we have: $\forall k \geq 0$,

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{g}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^k\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\|\mathbf{g}^{k+1} - \nabla\mathbf{f}(\mathbf{x}^{k+1})\|^2|\mathcal{F}^{k+1}\right]|\mathcal{F}^k\right] \\ &\leq 2L^2\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\|^2|\mathcal{F}^k\right] + 2nL^2\mathbb{E}\left[t^{k+1}|\mathcal{F}^k\right] \\ &\leq 2L^2\left(2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 2\alpha^2\mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2|\mathcal{F}^k\right]\right) + 2nL^2\left(2t^k + 3\alpha^2\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2 + \frac{9}{4mn}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right),\end{aligned}$$

where the second line uses (3.73) and the third line uses (3.71) and Corollary 3.3.2. The desired inequality then follows. □

3.3.5.5 Bounds on stochastic gradient tracking process

In this subsection, we analyze the variance-reduced stochastic gradient tracking process (3.69a). The analysis techniques presented here are related to [4,67], where uniformly bounded variance of each stochastic gradient is assumed as they focus on the online setting. In contrast, due to the local variance reduction technique in `GT-SAGA` that leverages the finite sum structure of the problem, such assumption is not needed here and we use Lemma 3.3.3 to control the variance of stochastic gradients.

Lemma 3.3.7. *The following inequality holds:* $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2\big] \leq \lambda^2 \mathbb{E}\big[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\big] + \lambda^2 \mathbb{E}\big[\|\mathbf{g}^{k+1} - \mathbf{g}^k\|^2\big] \\ + 2\mathbb{E}\big[\big\langle (\mathbf{W}-\mathbf{J})\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)\right)\big\rangle\big] \\ + 2\mathbb{E}\big[\big\langle (\mathbf{W}-\mathbf{J})\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k}) - \mathbf{g}^k\right)\big\rangle\big].\end{aligned}$$

Proof. Using (3.69a) and $\mathbf{J}\mathbf{W} = \mathbf{J}$, we have: $\forall k \geq 0$,

$$\begin{aligned}&\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2 \\ &= \|\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1} + (\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\|^2 \\ &= \|\mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2 + \|(\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\|^2 + 2\left\langle \mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\right\rangle \\ &\leq \lambda^2\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2 + \lambda^2\|\mathbf{g}^{k+1} - \mathbf{g}^k\|^2 + 2\left\langle \mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\right\rangle, \qquad (3.98)\end{aligned}$$

where the last line uses Lemma 3.3.1(a) and $\|\mathbf{W}-\mathbf{J}\| = \lambda$. To proceed, we observe that $\forall k \geq 0$,

$$\begin{aligned}&\mathbb{E}\left[\left\langle \mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\mathbf{g}^{k+1} - \mathbf{g}^k\right)\right\rangle | \mathcal{F}^{k+1}\right] \\ &= \left\langle \mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \mathbf{g}^k\right)\right\rangle \\ &= \left\langle \mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)\right)\right\rangle \\ &\quad + \left\langle \mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k}) - \mathbf{g}^k\right)\right\rangle, \qquad (3.99)\end{aligned}$$

where the first line uses that $\mathbb{E}[\mathbf{g}^{k+1}|\mathcal{F}^{k+1}] = \nabla \mathbf{f}(\mathbf{x}^{k+1})$ and that $\mathbf{y}^{k+1}$ and $\mathbf{g}^k$ are $\mathcal{F}^{k+1}$-measurable for all $k \geq 0$. We conclude by using (3.99) in (3.98) and taking the expectation. □

We next bound the third term in Lemma 3.3.7.

Lemma 3.3.8. *The following inequality holds:* $\forall k \geq 0$,

$$\begin{aligned}&\left\langle \mathbf{W}\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}, (\mathbf{W}-\mathbf{J})\left(\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)\right)\right\rangle \\ &\leq \left(\lambda\alpha L + 0.5\eta_1 + \eta_2\right)\lambda^2\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2 + 0.5\eta_1^{-1}\lambda^2\alpha^2L^2n\|\bar{\mathbf{g}}^k\|^2 + \eta_2^{-1}\lambda^2L^2\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2,\end{aligned}$$

where $\eta_1 > 0$ *and* $\eta_2 > 0$ *are arbitrary.*

Proof. Using Lemma 3.3.1(a) and $\|\mathbf{W}-\mathbf{J}\|=\lambda$, we have: $\forall k\geq 0$,

$$\left\langle \mathbf{W}\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1},(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}(\mathbf{x}^{k+1})-\nabla\mathbf{f}(\mathbf{x}^{k})\right)\right\rangle \leq \lambda^2 L\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|. \tag{3.100}$$

Observe that $\forall k\geq 0$,

$$\begin{aligned}\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\| &= \|\mathbf{x}^{k+1}-\mathbf{J}\mathbf{x}^{k+1}+\mathbf{J}\mathbf{x}^{k+1}-\mathbf{J}\mathbf{x}^{k}+\mathbf{J}\mathbf{x}^{k}-\mathbf{x}^{k}\|\\ &\leq \|\mathbf{x}^{k+1}-\mathbf{J}\mathbf{x}^{k+1}\|+\sqrt{n}\alpha\|\overline{\mathbf{g}}^{k}\|+\|\mathbf{x}^{k}-\mathbf{J}\mathbf{x}^{k}\|\\ &\leq 2\|\mathbf{x}^{k}-\mathbf{J}\mathbf{x}^{k}\|+\sqrt{n}\alpha\|\overline{\mathbf{g}}^{k}\|+\alpha\lambda\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|,\end{aligned} \tag{3.101}$$

where the last line is due to (3.72). We use (3.101) in (3.100) to obtain: $\forall k\geq 0$,

$$\begin{aligned}&\left\langle \mathbf{W}\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1},(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}(\mathbf{x}^{k+1})-\nabla\mathbf{f}(\mathbf{x}^{k})\right)\right\rangle\\ &\leq \lambda^3\alpha L\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2+\lambda^2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|\sqrt{n}\alpha L\|\overline{\mathbf{g}}^{k}\|+2\lambda^2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|L\|\mathbf{x}^{k}-\mathbf{J}\mathbf{x}^{k}\|.\end{aligned} \tag{3.102}$$

By Young's inequality, we have: $\forall k\geq 0$, for some $\eta_1>0$,

$$\lambda^2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|\sqrt{n}\alpha L\|\overline{\mathbf{g}}^{k}\|\leq 0.5\lambda^2\left(\eta_1\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2+\eta_1^{-1}n\alpha^2L^2\|\overline{\mathbf{g}}^{k}\|^2\right), \tag{3.103}$$

and, $\forall k\geq 0$, for some $\eta_2>0$,

$$2\lambda^2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|L\|\mathbf{x}^{k}-\mathbf{J}\mathbf{x}^{k}\|\leq \lambda^2\eta_2\|\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1}\|^2+\lambda^2\eta_2^{-1}L^2\|\mathbf{x}^{k}-\mathbf{J}\mathbf{x}^{k}\|^2. \tag{3.104}$$

The proof follows by applying (3.103) and (3.104) to (3.102). □

We next bound the fourth term in Lemma 3.3.7.

Lemma 3.3.9. *The following inequality holds:* $\forall k\geq 0$,

$$\mathbb{E}\left[\left\langle \mathbf{W}\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1},(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}(\mathbf{x}^{k})-\mathbf{g}^{k}\right)\right\rangle\right]\leq \mathbb{E}\left[\|\mathbf{g}^{k}-\nabla\mathbf{f}(\mathbf{x}^{k})\|^2\right]/n.$$

Proof. In the following, we denote $\nabla\mathbf{f}^k:=\nabla\mathbf{f}(\mathbf{x}^k)$ to simplify the notation. Observe that $\forall k\geq 0$,

$$\begin{aligned}&\mathbb{E}\left[\left\langle \mathbf{W}\mathbf{y}^{k+1}-\mathbf{J}\mathbf{y}^{k+1},(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}^{k}-\mathbf{g}^{k}\right)\right\rangle|\mathcal{F}^k\right]\\ &\overset{(i)}{=}\mathbb{E}\left[\left\langle \mathbf{W}^2(\mathbf{y}^{k}+\mathbf{g}^{k}-\mathbf{g}^{k-1}),(\mathbf{W}-\mathbf{J})(\nabla\mathbf{f}^{k}-\mathbf{g}^{k})\right\rangle|\mathcal{F}^k\right]\\ &\overset{(ii)}{=}\mathbb{E}\left[\left\langle \mathbf{W}^2\mathbf{g}^{k},(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}^{k}-\mathbf{g}^{k}\right)\right\rangle|\mathcal{F}^k\right]\\ &\overset{(iii)}{=}\mathbb{E}\left[\left\langle \mathbf{W}^2\left(\mathbf{g}^{k}-\nabla\mathbf{f}^{k}\right),(\mathbf{W}-\mathbf{J})\left(\nabla\mathbf{f}^{k}-\mathbf{g}^{k}\right)\right\rangle|\mathcal{F}^k\right]\\ &\overset{(iv)}{=}\mathbb{E}\left[(\mathbf{g}^{k}-\nabla\mathbf{f}^{k})^\top(\mathbf{J}-\mathbf{W}^\top\mathbf{W}^2)(\mathbf{g}^{k}-\nabla\mathbf{f}^{k})|\mathcal{F}^k\right],\end{aligned} \tag{3.105}$$

where (i) uses (3.69a) and $\mathbf{J}\mathbf{W}=\mathbf{J}$, (ii) and (iii) use that $\mathbf{y}^k$, $\mathbf{g}^{k-1}$ and $\nabla\mathbf{f}^k$ are $\mathcal{F}^k$-measurable and that $\mathbb{E}[\mathbf{g}^k|\mathcal{F}^k]=\nabla\mathbf{f}^k$ for all $k\geq 0$, and (iv) uses $\mathbf{J}\mathbf{W}=\mathbf{J}$. Using

$$\mathbb{E}\left[\left\langle \mathbf{g}_i^k-\nabla f_i(\mathbf{x}_i^k),\mathbf{g}_j^k-\nabla f_j(\mathbf{x}_j^k)\right\rangle|\mathcal{F}^k\right]=0$$

for all $i \neq j \in \mathcal{V}$ and the fact that $\mathbf{W}^\top \mathbf{W}^2$ is nonnegative, we have: $\forall k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[(\mathbf{g}^k - \nabla \mathbf{f}^k)^\top \left(\mathbf{J} - \mathbf{W}^\top \mathbf{W}^2\right)\left(\mathbf{g}^k - \nabla \mathbf{f}^k\right) | \mathcal{F}^k\right] \\
&= \mathbb{E}\left[(\mathbf{g}^k - \nabla \mathbf{f}^k)^\top \mathrm{diag}\left(\mathbf{J} - \mathbf{W}^\top \mathbf{W}^2\right)\left(\mathbf{g}^k - \nabla \mathbf{f}^k\right) | \mathcal{F}^k\right] \\
&\leq \mathbb{E}\left[(\mathbf{g}^k - \nabla \mathbf{f}^k)^\top \mathrm{diag}(\mathbf{J})\left(\mathbf{g}^k - \nabla \mathbf{f}^k\right) | \mathcal{F}^k\right]. \qquad (3.106)
\end{aligned}$$

The proof follows by taking the expectation of (3.106). □

We finally bound the second term in Lemma 3.3.7.

Lemma 3.3.10. *The following inequality holds:* $\forall k \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{g}^{k+1} - \mathbf{g}^k\|^2\right] \leq\ & 12\lambda^2\alpha^2 L^2 \mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\right] + 2\mathbb{E}\left[\|\mathbf{g}^k - \nabla \mathbf{f}(\mathbf{x}^k)\|^2\right] + \mathbb{E}\left[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2\right] \\
&+ 18L^2 \mathbb{E}\left[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] + 6n\alpha^2 L^2 \mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2\right].
\end{aligned}$$

Proof. Since $\mathbf{g}^k$ and $\nabla \mathbf{f}(\mathbf{x}^{k+1})$ are $\mathcal{F}^{k+1}$-measurable, and $\mathbb{E}[\mathbf{g}^{k+1}|\mathcal{F}^{k+1}] = \nabla \mathbf{f}(\mathbf{x}^{k+1})$, we have: $\forall k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{g}^{k+1} - \mathbf{g}^k\|^2 | \mathcal{F}^{k+1}\right] \\
&= \mathbb{E}\left[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2 | \mathcal{F}^{k+1}\right] + \|\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \mathbf{g}^k\|^2 \\
&\leq \mathbb{E}\left[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2 | \mathcal{F}^{k+1}\right] + 2\|\nabla \mathbf{f}(\mathbf{x}^{k+1}) - \nabla \mathbf{f}(\mathbf{x}^k)\|^2 + 2\|\nabla \mathbf{f}(\mathbf{x}^k) - \mathbf{g}^k\|^2 \\
&\leq \mathbb{E}\left[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2 | \mathcal{F}^{k+1}\right] + 2L^2\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + 2\|\nabla \mathbf{f}(\mathbf{x}^k) - \mathbf{g}^k\|^2. \qquad (3.107)
\end{aligned}$$

Similar to the derivation of (3.101), we have: $\forall k \geq 0$,

$$\begin{aligned}
\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 &= \|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1} + \mathbf{J}\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^k + \mathbf{J}\mathbf{x}^k - \mathbf{x}^k\|^2 \\
&\leq 3\|\mathbf{x}^{k+1} - \mathbf{J}\mathbf{x}^{k+1}\|^2 + 3n\alpha^2\|\overline{\mathbf{g}}^k\|^2 + 3\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 \\
&\leq 9\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + 3n\alpha^2\|\overline{\mathbf{g}}^k\|^2 + 6\alpha^2\lambda^2\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2,
\end{aligned}$$

where the last line is due to (3.71). We conclude by applying the last line above to (3.107). □

Now, we apply Lemma 3.3.8, 3.3.9, 3.3.10 to Lemma 3.3.7.

Lemma 3.3.11. *The following inequality holds:* $\forall k \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2\right] \leq\ & \left(1 + 2\lambda\alpha L + \eta_1 + 2\eta_2 + 12\lambda^2\alpha^2 L^2\right)\lambda^2 \mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\right] \\
&+ \left(2\eta_2^{-1} + 18\right)\lambda^2 L^2 \mathbb{E}\left[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] + \left(\eta_1^{-1} + 6\right)\lambda^2\alpha^2 L^2 n \mathbb{E}\left[\|\overline{\mathbf{g}}^k\|^2\right] \\
&+ \left(2\lambda^2 + 2/n\right)\mathbb{E}\left[\|\mathbf{g}^k - \nabla \mathbf{f}(\mathbf{x}^k)\|^2\right] + \lambda^2 \mathbb{E}\left[\|\mathbf{g}^{k+1} - \nabla \mathbf{f}(\mathbf{x}^{k+1})\|^2\right].
\end{aligned}$$

Proof. Apply Lemma 3.3.8, 3.3.9, 3.3.10 to Lemma 3.3.7. □

Finally, we use Lemma 3.3.3 and 3.3.6 to refine Lemma 3.3.11 and establish a contraction in the gradient tracking process.

Lemma 3.3.12. *If* $0 < \alpha \leq \min\left\{\frac{1-\lambda^2}{16\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\right\}\frac{1}{L}$, *then we have:* $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2\big] \leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\big] + \frac{30.5L^2}{1-\lambda^2}\mathbb{E}\big[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\big] \\ + \frac{97L^2 n}{8}\mathbb{E}\big[t^k\big] + \frac{16\lambda^2\alpha^2L^2n}{1-\lambda^2}\mathbb{E}\big[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\big].\end{aligned}$$

Proof. We apply Lemma 3.3.3 and 3.3.6 to Lemma 3.3.11 to obtain: if $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, then $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2\right] \leq& \left(1 + 2\lambda\alpha L + \eta_1 + 2\eta_2 + \left(12\lambda^2 + 4\right)\alpha^2L^2\right)\lambda^2\mathbb{E}\left[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\right] \\ &+\left((2\eta_2^{-1} + 18)\lambda^2 + (\eta_1^{-1} + 6)\frac{2\lambda^2\alpha^2L^2}{n} + \frac{4}{n} + 12.5\lambda^2\right)L^2\mathbb{E}\left[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \\ &+\left(2(\eta_1^{-1} + 6)\lambda^2\alpha^2L^2 + \left(2\lambda^2 + 1/n\right)4n\right)L^2\mathbb{E}\big[t^k\big] \\ &+(\eta_1^{-1} + 12)\lambda^2\alpha^2L^2n\mathbb{E}\left[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\right]. \qquad (3.108)\end{aligned}$$

We fix $\eta_1 = \frac{1-\lambda^2}{16\lambda^2}$ and $\eta_2 = \frac{1-\lambda^2}{8\lambda^2}$. It can then be verified that $1 + 2\lambda\alpha L + \eta_1 + 2\eta_2 + (12\lambda^2 + 4)\alpha^2L^2 \leq \frac{1+\lambda^2}{2\lambda^2}$, if $0 < \alpha \leq \frac{1-\lambda^2}{16\lambda L}$. The proof then follows by applying this inequality and the values of η_1 and η_2 to (3.108). □

3.3.5.6 Proof of Theorem 3.3.1

In this subsection, we prove the convergence of **GT-SAGA** for general smooth non-convex functions. To this aim, we write the contraction inequalities in (3.70), Corollary 3.3.1, and Lemma 3.3.12 as a linear time-invariant (LTI) dynamics that jointly characterizes the evolution of the consensus, gradient tracking, and the auxiliary sequence t^k.

Proposition 3.3.1. *If* $0 < \alpha \leq \min\left\{\frac{1-\lambda^2}{16\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\right\}\frac{1}{L}$, *then*

$$\mathbf{u}^{k+1} \leq \mathbf{G}_\alpha\mathbf{u}^k + \mathbf{b}^k, \qquad \forall k \geq 0,$$

where $\mathbf{u}^k \in \mathbb{R}^3$, $\mathbf{G}_\alpha \in \mathbb{R}^{3\times 3}$, *and* $\mathbf{b}^k \in \mathbb{R}^3$ *are given by*

$$\mathbf{u}^k := \begin{bmatrix}\mathbb{E}\left[\frac{\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2}{n}\right] \\ \mathbb{E}\left[t^k\right] \\ \mathbb{E}\left[\frac{\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2}{nL^2}\right]\end{bmatrix}, \qquad \mathbf{b} := \begin{bmatrix}0 \\ 4m\alpha^2 \\ \frac{16\lambda^2\alpha^2}{1-\lambda^2}\end{bmatrix}, \qquad \mathbf{G}_\alpha := \begin{bmatrix}\frac{1+\lambda^2}{2} & 0 & \frac{2\lambda^2\alpha^2L^2}{1-\lambda^2} \\ \frac{9}{4m} & 1 - \frac{1}{4m} & 0 \\ \frac{30.5}{1-\lambda^2} & \frac{97}{8} & \frac{1+\lambda^2}{2}\end{bmatrix},$$

and $\mathbf{b}^k := \mathbf{b}\mathbb{E}\left[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\right]$.

We first derive the range of the step-size α under which the spectral radius of $\mathbf{G}_\alpha$ defined in Proposition 3.3.1 is less than 1, with the help of the following Lemma from [36].

Lemma 3.3.13. *Let $\mathbf{X} \in \mathbb{R}^{d\times d}$ be a non-negative matrix and $\mathbf{x} \in \mathbb{R}^d$ be a positive vector. If $\mathbf{X}\mathbf{x} < \mathbf{x}$, then $\rho(\mathbf{X}) < 1$. Moreover, if $\mathbf{X}\mathbf{x} \leq \beta\mathbf{x}$, for some $\beta \in \mathbb{R}$, then $\rho(\mathbf{X}) \leq \beta$.*

Lemma 3.3.14. *If $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{35\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\right\}\frac{1}{L}$, then $\rho(\mathbf{G}_\alpha) < 1$ and thus $\sum_{k=0}^{\infty}\mathbf{G}_\alpha^k = (\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}$.*

Proof. In light of Lemma 3.3.13, we find a positive vector $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \epsilon_3]^\top$ and the range of α such that $\mathbf{G}_\alpha\boldsymbol{\epsilon} < \boldsymbol{\epsilon}$, which is equivalent to the following set of inequalities:

$$\alpha^2 < \frac{(1-\lambda^2)^2}{4\lambda^2L^2}\frac{\epsilon_1}{\epsilon_3}, \tag{3.109}$$

$$9\epsilon_1 < \epsilon_2, \tag{3.110}$$

$$\frac{61}{(1-\lambda^2)^2}\epsilon_1 + \frac{97}{4(1-\lambda^2)}\epsilon_2 < \epsilon_3, \tag{3.111}$$

Based on (3.110), we set $\epsilon_1 = 1$ and $\epsilon_2 = 10$. Then based on (3.111), we set $\epsilon_3 = \frac{303.5}{(1-\lambda^2)^2}$. The proof follows by using the values of ϵ_1 and ϵ_3 in (3.109). □

Based on the LTI dynamics in Proposition 3.3.1, we derive the following lemma that is the key to establish the convergence of `GT-SAGA` for general smooth nonconvex functions.

Lemma 3.3.15. *If $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{35\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\right\}\frac{1}{L}$, then we have: $\forall K \geq 1$,*

$$\sum_{k=0}^{K}\mathbf{u}^k \leq (\mathbf{I} - \mathbf{G}_\alpha)^{-1}\left(\mathbf{u}^0 + \mathbf{b}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\big]\right).$$

Proof. We recursively apply the dynamics in Proposition 3.3.1 to obtain: $\mathbf{u}^k \leq \mathbf{G}_\alpha^k\mathbf{u}^0 + \sum_{r=0}^{k-1}\mathbf{G}_\alpha^r\mathbf{b}^{k-1-r}, \forall k \geq 1$. We sum this inequality over k to obtain: $\forall K \geq 1$,

$$\begin{aligned}\sum_{k=0}^{K}\mathbf{u}^k &\leq \sum_{k=0}^{K}\mathbf{G}_\alpha^k\mathbf{u}^0 + \sum_{k=1}^{K}\sum_{r=0}^{k-1}\mathbf{G}_\alpha^r\mathbf{b}^{k-1-r}\\ &\leq \left(\sum_{k=0}^{\infty}\mathbf{G}_\alpha^k\right)\mathbf{u}^0 + \sum_{k=0}^{K-1}\left(\sum_{k=0}^{\infty}\mathbf{G}_\alpha^k\right)\mathbf{b}^k.\end{aligned}$$

The proof follows by $\sum_{k=0}^{\infty}\mathbf{G}_\alpha^k = (\mathbf{I} - \mathbf{G}_\alpha)^{-1}$ and the definition of $\mathbf{b}^k$ in Proposition 3.3.1. □

Lemma 3.3.16. *If $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\right\}\frac{1}{L}$, then*

$$(\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1} \leq \begin{bmatrix}\star & \frac{776\lambda^2m\alpha^2L^2}{(1-\lambda^2)^3} & \frac{16\lambda^2\alpha^2L^2}{(1-\lambda^2)^3}\\ \star & 8m & \frac{114\lambda^2\alpha^2L^2}{(1-\lambda^2)^3}\\ \star & \star & \star\end{bmatrix}, \qquad (\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}\mathbf{b} \leq \begin{bmatrix}\left(3104m^2 + \frac{256\lambda^2}{1-\lambda^2}\right)\frac{\lambda^2\alpha^4L^2}{(1-\lambda^2)^3}\\ 33m^2\alpha^2\\ \star\end{bmatrix},$$

where the $\star$ entries are not needed for further derivations.

Proof. In the following, for a matrix $\mathbf{X}$, we denote $\mathbf{X}^*$ as its adjugate and $[\mathbf{X}]_{i,j}$ as its (i,j)-th entry. We first note that if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{48\lambda L}$, $\det(\mathbf{I}_3 - \mathbf{G}_\alpha) \geq \frac{(1-\lambda^2)^2}{32m}$. We next derive upper bounds for entries of $(\mathbf{I}_3 - \mathbf{G}_\alpha)^*$:

$$[(\mathbf{I}_3 - \mathbf{G}_\alpha)^*]_{1,2} = \frac{97\lambda^2\alpha^2L^2}{4\,(1-\lambda^2)}, \qquad [(\mathbf{I}_3 - \mathbf{G}_\alpha)^*]_{1,3} = \frac{\lambda^2\alpha^2L^2}{2m(1-\lambda^2)},$$

$$[(\mathbf{I}_3 - \mathbf{G}_\alpha)^*]_{2,2} \leq \frac{(1-\lambda^2)^2}{4}, \qquad [(\mathbf{I}_3 - \mathbf{G}_\alpha)^*]_{2,3} = \frac{9\lambda^2\alpha^2L^2}{2m(1-\lambda^2)}.$$

The upper bound on $(\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}$ then follows by using the above relations. Finally, we have:

$$(\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}\mathbf{b} \leq \begin{bmatrix} \frac{3104\lambda^2m^2\alpha^4L^2}{(1-\lambda^2)^3} + \frac{256\lambda^4\alpha^4L^2}{(1-\lambda^2)^4} \\ 32m^2\alpha^2 + \frac{2304\lambda^4\alpha^4L^2}{(1-\lambda^2)^4} \\ \star \end{bmatrix}.$$

If $0 < \alpha \leq \frac{(1-\lambda^2)^2}{48\lambda L}$, then $32m^2\alpha^2 + \frac{2304\lambda^4\alpha^4L^2}{(1-\lambda^2)^4} \leq 33m^2\alpha^2$ and the bound on $(\mathbf{I}_3 - \mathbf{G}_\alpha)^{-1}\mathbf{b}$ follows. □

We now bound two important quantities as follows.

Lemma 3.3.17. *If* $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\right\}\frac{1}{L}$, *then we have:* $\forall K \geq 1$,

$$\sum_{k=0}^{K}\mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \leq \frac{16\lambda^4\alpha^2}{(1-\lambda^2)^3}\frac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n} + \left(97m^2 + \frac{8\lambda^2}{1-\lambda^2}\right)\frac{32\lambda^2\alpha^4L^2}{(1-\lambda^2)^3}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\big], \quad (3.112)$$

$$\sum_{k=0}^{K}\mathbb{E}\left[t^k\right] \leq \frac{114\lambda^4\alpha^2}{(1-\lambda^2)^3}\frac{\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2}{n} + 33m^2\alpha^2\sum_{k=0}^{K-1}\mathbb{E}\big[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\big]. \quad (3.113)$$

Proof. By (3.69a), we have $\|\mathbf{y}^1 - \mathbf{J}\mathbf{y}^1\|^2 = \|(\mathbf{W} - \mathbf{J})(\mathbf{y}^0 + \mathbf{g}^0 - \mathbf{g}^{-1})\|^2 \leq \lambda^2\|\nabla\mathbf{f}(\mathbf{x}^0)\|^2$. The proof then follows by applying this inequality and Lemma 3.3.16 to Lemma 3.3.15. □

Now, we are ready to prove Theorem 3.3.1.

Proof of Theorem 3.3.1. We sum up the inequality in Lemma 3.3.4 over k: if $0 < \alpha \leq \frac{1}{2L}$, then $\forall K \geq 1$,

$$\mathbb{E}\big[F(\overline{\mathbf{x}}^K)\big] \leq F(\overline{\mathbf{x}}^0) - \frac{\alpha}{2}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla F(\overline{\mathbf{x}}^k)\|^2\big] - \frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\big] + \frac{\alpha^2L^3}{n}\sum_{k=0}^{K-1}\mathbb{E}\big[t^k\big] + \alpha L^2\sum_{k=0}^{K-1}\mathbb{E}\Big[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\Big]. \quad (3.114)$$

By the L-smoothness of F, we have: $\frac{1}{2n}\sum_{i=1}^{n}\|\nabla F(\mathbf{x}_i^k)\|^2 \leq \|\nabla F(\overline{\mathbf{x}}^k)\|^2 + \frac{L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2$, $\forall k \geq 0$. Using this inequality in (3.114), we obtain: if $0 < \alpha \leq \frac{1}{2L}$, then $\forall K \geq 1$,

$$\mathbb{E}\big[F(\overline{\mathbf{x}}^K)\big] \leq F(\overline{\mathbf{x}}^0) - \frac{\alpha}{4n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla F(\mathbf{x}_i^k)\|^2\big] - \frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\overline{\nabla\mathbf{f}}(\mathbf{x}^k)\|^2\big] + \frac{\alpha^2L^3}{n}\sum_{k=0}^{K-1}\mathbb{E}\big[t^k\big] + \frac{3\alpha L^2}{2}\sum_{k=0}^{K-1}\mathbb{E}\Big[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\Big]. \quad (3.115)$$

Applying (3.113) to (3.115), we obtain the following: if $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}, \frac{1}{2}\right\}\frac{1}{L}$, then $\forall K \geq 1$,

$$\mathbb{E}\big[F(\overline{\mathbf{x}}^K)\big] \leq F(\overline{\mathbf{x}}^0) - \frac{\alpha}{4n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla F(\mathbf{x}_i^k)\|^2\big] - \frac{\alpha}{8}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2\big] + \frac{114\lambda^4\alpha^4L^3}{n(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n}$$
$$+ \frac{3\alpha L^2}{2}\sum_{k=0}^{K-1}\mathbb{E}\Big[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\Big] - \frac{\alpha}{8}\left(1 - \frac{264m^2\alpha^3L^3}{n}\right)\sum_{k=0}^{K-1}\mathbb{E}\big[\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2\big]. \tag{3.116}$$

If $0 < \alpha \leq \frac{2n^{1/3}}{13m^{2/3}L}$, $1 - \frac{264m^2\alpha^3L^3}{n} \geq 0$ and thus the last term in (3.116) may be dropped. We then use (3.112) in (3.116) to obtain: if $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{48\lambda}, \frac{2n^{1/3}}{13m^{2/3}}, \frac{1}{2}\right\}\frac{1}{L}$, then $\forall K \geq 1$,

$$\mathbb{E}\big[F(\overline{\mathbf{x}}^K)\big] \leq F(\overline{\mathbf{x}}^0) - \frac{\alpha}{4n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla F(\mathbf{x}_i^k)\|^2\big] - \frac{\alpha L^2}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right]$$
$$+ \left(\frac{114\alpha L}{28n} + 1\right)\frac{28\lambda^4\alpha^3L^2}{(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n}$$
$$- \frac{\alpha}{8}\left(1 - \max\left\{97m^2, \frac{8\lambda^2}{1-\lambda^2}\right\}\frac{896\lambda^2\alpha^4L^4}{(1-\lambda^2)^3}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2\right]. \tag{3.117}$$

If $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^{3/4}}{18\lambda^{1/2}m^{1/2}}, \frac{1-\lambda^2}{12\lambda}\right\}\frac{1}{L}$, then $\max\left\{97m^2, \frac{8\lambda^2}{1-\lambda^2}\right\}\frac{896\lambda^2\alpha^4L^4}{(1-\lambda^2)^3} \leq 1$ and the last term in (3.117) may be dropped. Thus, if $0 < \alpha \leq \overline{\alpha}_1$ for $\overline{\alpha}_1$ defined in Theorem 3.3.1, we obtain from (3.117) that

$$\mathbb{E}\big[F(\overline{\mathbf{x}}^K)\big] \leq F(\overline{\mathbf{x}}^0) - \frac{\alpha}{4n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla F(\mathbf{x}_i^k)\|^2\big] - \frac{\alpha L^2}{4}\sum_{k=0}^{K-1}\mathbb{E}\Big[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\Big] + \frac{112\lambda^4\alpha^3L^2}{(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n}.$$

for all $K \geq 1$. Since F is bounded below by F^*, the above inequality leads to, $\forall K \geq 1$,

$$\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_i^k)\|^2 + L^2\|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2\Big] \leq \frac{4(F(\overline{\mathbf{x}}^0) - F^*)}{\alpha} + \frac{448\lambda^4\alpha^2L^2}{(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n}. \tag{3.118}$$

Since the RHS of (3.118) is finite and independent of K, we let $K \to \infty$ in (3.118) to obtain:

$$\sum_{k=0}^{\infty}\sum_{i=1}^{n}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_i^k)\|^2 + \|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2\Big] < \infty, \tag{3.119}$$

which shows that all nodes in **GT-SAGA** asymptotically agree on a stationary point of F in the mean-squared sense. Moreover, since the series on the LHS of (3.119) is nonnegative, we may exchange the order of the series and expectation to obtain [127]: $\mathbb{E}\left[\sum_{k=0}^{\infty}\sum_{i=1}^{n}(\|\nabla F(\mathbf{x}_i^k)\|^2 + \|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2)\right] < \infty$, which implies that

$$\mathbb{P}\left(\sum_{k=0}^{\infty}\sum_{i=1}^{n}\left(\|\nabla F(\mathbf{x}_i^k)\|^2 + \|\mathbf{x}_i^k - \overline{\mathbf{x}}^k\|^2\right) < \infty\right) = 1, \tag{3.120}$$

i.e., all nodes in **GT-SAGA** asymptotically agree on a stationary point of F in the almost sure sense. Finally, towards the iteration complexity of **GT-SAGA**, we set $\alpha = \overline{\alpha}_1$ in (3.118) and divide the resulting inequality by K to obtain: $\forall K \geq 1$,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla F(\mathbf{x}_i^k)\|^2\big] \leq \frac{4(F(\overline{\mathbf{x}}^0) - F^*)}{\overline{\alpha}_1 K} + \frac{448\lambda^4\overline{\alpha}_1^2L^2}{(1-\lambda^2)^3K}\frac{\|\nabla \mathbf{f}(\mathbf{x}^0)\|^2}{n}. \tag{3.121}$$

Based on (3.121), the iteration complexity of **GT-SAGA** then follows by recalling the definition of $\overline{\alpha}_1$ in Theorem 3.3.1 and that $\frac{448\lambda^4\overline{\alpha}_1^2L^2}{(1-\lambda^2)^3} \leq \frac{\lambda^2(1-\lambda^2)}{4}$ since $0 < \overline{\alpha}_1 \leq \frac{(1-\lambda^2)^2}{48\lambda L}$. □

3.3.5.7 Proof of Theorem 3.3.2

In this subsection, we prove the linear rate of **GT-SAGA** when the global function F additionally satisfies the PL condition. In particular, we use the PL condition and Lemma 3.3.1(e) to refine the descent inequality in Lemma 3.3.4 and the previously obtained LTI system in Proposition 3.3.1.

Lemma 3.3.18. *If* $0 < \alpha \leq \frac{1}{2L}$, *then* $\forall k \geq 0$,

$$\mathbb{E}\big[F(\overline{\mathbf{x}}^{k+1}) - F^* | \mathcal{F}^k\big] \leq (1-\mu\alpha)(F(\overline{\mathbf{x}}^k) - F^*) + \frac{\alpha L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2 + \frac{\alpha^2 L^3}{n} t^k.$$

Proof. Apply the PL condition to Lemma 3.3.4 and then subtract F^* from the resulting inequality. □

Next, we refine Corollary 3.3.1 as follows.

Lemma 3.3.19. *If* $0 < \alpha \leq \frac{\sqrt{n}}{\sqrt{8m}L}$, *then* $\forall k \geq 0$,

$$\mathbb{E}\big[t^{k+1} | \mathcal{F}^k\big] \leq \Big(1 - \frac{1}{4m}\Big) t^k + 16m\alpha^2 L \left(F(\overline{\mathbf{x}}^k) - F^*\right) + \Big(8m\alpha^2 L^2 + \frac{9}{4m}\Big)\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2.$$

Proof. By Lemma 3.3.1(c) and 3.3.1(e), we have: $\forall k \geq 0$,

$$\begin{aligned}\|\overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 &\leq 2\|\nabla F(\overline{\mathbf{x}}^k)\|^2 + 2\|\nabla F(\overline{\mathbf{x}}^k) - \overline{\nabla \mathbf{f}}(\mathbf{x}^k)\|^2 \\ &\leq 4L\left(F(\overline{\mathbf{x}}^k) - F^*\right) + \frac{2L^2}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2. \end{aligned} \tag{3.122}$$

The proof follows by applying (3.122) to Corollary 3.3.1. □

We finally refine Lemma 3.3.12 as follows.

Lemma 3.3.20. *If* $0 < \alpha \leq \min\left\{\frac{1-\lambda^2}{16\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\right\}\frac{1}{L}$, *then* $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2\big] \leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\big] + \frac{31L^2}{1-\lambda^2}\mathbb{E}\big[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\big] \\ + \frac{97L^2 n}{8}\mathbb{E}\big[t^k\big] + \frac{64\lambda^2\alpha^2 L^3 n}{1-\lambda^2}\mathbb{E}\big[F(\overline{\mathbf{x}}^k) - F^*\big].\end{aligned}$$

Proof. Applying (3.122) to Lemma 3.3.12, we have: if $0 < \alpha \leq \min\left\{\frac{1-\lambda^2}{16\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}\right\}\frac{1}{L}$, then $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{y}^{k+2} - \mathbf{J}\mathbf{y}^{k+2}\|^2\big] \leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\big] + \big(30.5 + 32\lambda^2\alpha^2 L^2\big)\frac{L^2}{1-\lambda^2}\mathbb{E}\big[\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\big] \\ + \frac{97L^2 n}{8}\mathbb{E}\big[t^k\big] + \frac{64\lambda^2\alpha^2 L^3 n}{1-\lambda^2}\mathbb{E}\big[F(\overline{\mathbf{x}}^k) - F^*\big].\end{aligned}$$

We conclude by $30.5 + 32\lambda^2\alpha^2 L^2 \leq 31$ if $0 < \alpha \leq \frac{1-\lambda^2}{16\lambda L}$. □

Now, we write (3.70), Lemma 3.3.18, 3.3.19 and 3.3.20 in a LTI system.

Proposition 3.3.2. *If* $0 < \alpha \leq \min\left\{\frac{1-\lambda^2}{16\lambda}, \frac{\sqrt{n}}{\sqrt{8m}}, \frac{1}{2}\right\}\frac{1}{L}$, *then*

$$\mathbf{v}^{k+1} \leq \mathbf{H}_\alpha \mathbf{v}^k, \qquad \forall k \geq 0,$$

where $\mathbf{v}^k \in \mathbb{R}^4$ *and* $\mathbf{H}_\alpha \in \mathbb{R}^{4\times 4}$ *are given by*

$$\mathbf{v}^k := \begin{bmatrix} \mathbb{E}\left[\frac{1}{n}\|\mathbf{x}^k - \mathbf{J}\mathbf{x}^k\|^2\right] \\ \frac{1}{L}\mathbb{E}\left[F(\overline{\mathbf{x}}^k) - F^*\right] \\ \mathbb{E}\left[t^k\right] \\ \mathbb{E}\left[\frac{1}{nL^2}\|\mathbf{y}^{k+1} - \mathbf{J}\mathbf{y}^{k+1}\|^2\right] \end{bmatrix}, \qquad \mathbf{H}_\alpha := \begin{bmatrix} \frac{1+\lambda^2}{2} & 0 & 0 & \frac{2\lambda^2\alpha^2L^2}{1-\lambda^2} \\ \alpha L & 1-\mu\alpha & \frac{\alpha^2L^2}{n} & 0 \\ 8m\alpha^2L^2 + \frac{9}{4m} & 16m\alpha^2L^2 & 1-\frac{1}{4m} & 0 \\ \frac{31}{1-\lambda^2} & \frac{64\lambda^2\alpha^2L^2}{1-\lambda^2} & \frac{97}{8} & \frac{1+\lambda^2}{2} \end{bmatrix}.$$

We are ready to prove Theorem 3.3.2, i.e., to establish an upper bound on $\rho(\mathcal{H}_\alpha)$ that characterizes the explicit linear rate of `GT-SAGA` under the PL condition.

Proof of Theorem 3.3.2. In light of Lemma 3.3.13, we solve for the range of α under which there exists a positive vector $\mathbf{s}_\alpha = [s_1, s_2, s_3, s_4]^\top$ s.t. $\mathbf{H}_\alpha \mathbf{s}_\alpha \leq (1 - \frac{\mu\alpha}{2})\mathbf{s}_\alpha$, i.e.,

$$\frac{2\lambda^2\alpha^2L^2}{1-\lambda^2}s_4 \leq \left(\frac{1-\lambda^2}{2} - \frac{\mu\alpha}{2}\right)s_1, \tag{3.123}$$

$$\alpha L s_1 + \frac{\alpha^2L^2}{n}s_3 \leq \frac{\mu\alpha}{2}s_2, \tag{3.124}$$

$$\left(8m\alpha^2L^2 + \frac{9}{4m}\right)s_1 + 16m\alpha^2L^2 s_2 \leq \frac{1-2m\mu\alpha}{4m}s_3, \tag{3.125}$$

$$\frac{31}{1-\lambda^2}s_1 + \frac{64\lambda^2\alpha^2L^2}{1-\lambda^2}s_2 + \frac{97}{8}s_3 \leq \frac{1-\lambda^2-\mu\alpha}{2}s_4. \tag{3.126}$$

We first note that (3.124) is equivalent to $\frac{\alpha L^2}{n}s_3 \leq \frac{\mu}{2}s_2 - Ls_1$ and hence we set the values of s_1, s_2, s_3 as

$$s_1 = 1/(4\kappa), \qquad s_2 = 1, \qquad s_3 = n/(4\alpha\kappa L), \tag{3.127}$$

where $\kappa = L/\mu$. Next, we write (3.125) equivalently as

$$8m\alpha^2L^2\left(s_1 + 2s_2\right) \leq \frac{1-2m\mu\alpha}{4m}s_3 - \frac{9}{4m}s_1. \tag{3.128}$$

According to (3.128), we enforce $0 < \alpha \leq \frac{1}{4m\mu}$, i.e., $\frac{1-2m\mu\alpha}{4m} \geq \frac{1}{8m}$; therefore to make (3.128) hold, with the help of the values of s_1, s_2, s_3 in (3.127), it suffices to further choose α such that

$$18m\alpha^2L^2 \leq \frac{1}{16m\kappa}\left(\frac{n}{2\alpha L} - 9\right). \tag{3.129}$$

According to (3.129), we enforce $0 < \alpha \leq \frac{n}{36L}$, i.e., $\frac{n}{2\alpha L} - 9 \geq \frac{n}{4\alpha L}$, and therefore to make (3.129) hold, it suffices to further choose α such that $0 < \alpha \leq \frac{n^{1/3}}{10.5m^{2/3}\kappa^{1/3}L}$. Next, according to (3.126) we further enforce $0 < \alpha \leq \frac{1-\lambda^2}{2\mu}$, i.e., $\frac{1-\lambda^2-\mu\alpha}{2} \geq \frac{1-\lambda^2}{4}$ and therefore to make (3.126) hold we set s_4 as

$$s_4 = \frac{124}{(1-\lambda^2)^2}s_1 + \frac{256\lambda^2\alpha^2L^2}{(1-\lambda^2)^2}s_2 + \frac{97}{2(1-\lambda^2)}s_3. \tag{3.130}$$

Finally, since $0 < \alpha \leq \frac{1-\lambda^2}{2\mu}$, to make (3.123) hold, it suffices to further choose α such that $\frac{8\lambda^2\alpha^2L^2}{(1-\lambda^2)^2}\frac{s_4}{s_1} \leq 1$, which, using the values of s_1, s_4, becomes

$$\frac{992\lambda^2\alpha^2L^2}{(1-\lambda^2)^4} + \frac{8192\kappa\lambda^4\alpha^4L^4}{(1-\lambda^2)^4} + \frac{388\lambda^2 n\alpha L}{(1-\lambda^2)^3} \leq 1. \tag{3.131}$$

If $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{55\lambda}, \frac{1-\lambda^2}{13\lambda\kappa^{1/4}}, \frac{(1-\lambda^2)^3}{388\lambda^2 n}\right\}\frac{1}{L}$, then the terms on the LHS of (3.131) are respectively less than $\frac{1}{3}$ and thus (3.131) holds. Based on the above derivations and Lemma 3.3.13, we have: if $0 < \alpha \leq \overline{\alpha}_2$ for $\overline{\alpha}_2$ defined in Theorem 3.3.2, then $\rho(\mathbf{H}_\alpha) \leq 1 - \frac{\mu\alpha}{2}$ which concludes the proof. □

3.4 Conclusion

In this chapter, we consider decentralized empirical risk minimization problems defined in a network of n nodes, where each node holds m smooth non-convex cost functions. The goal of the networked nodes is to find an ϵ-accurate first-order stationary point of the average of $N := nm$ cost functions across all nodes. For this formulation, we consider two instances of the **GT-VR** framework proposed in Chapter 2, called **GT-SARAH** and **GT-SAGA**, that exhibit trade-offs in regimes of practical interest. In a big-data regime $n = \mathcal{O}(N^{1/2}(1-\lambda)^3)$, the gradient complexity of **GT-SARAH** reduces to $\mathcal{O}(LN^{1/2}\epsilon^{-2})$ which matches that of the centralized optimal methods [48–50] for this problem class, where L is the smoothness parameter of the cost functions and $(1-\lambda)$ is the spectral gap of the network weight matrix. On the other hand, in large-scale network regimes where the number of the nodes and the spectral gap of the network are large, we show that **GT-SAGA** achieves faster convergence than **GT-SARAH** and other existing approaches.

Chapter 4

Decentralized Online Stochastic Non-Convex Optimization

In this chapter, we study decentralized smooth non-convex expected risk minimization problems. In particular, we establish the convergence properties of the well-known `GT-DSGD` algorithm which combines `DSGD` with the gradient tracking technique. For general smooth non-convex functions, we establish the conditions under which `GT-DSGD` exhibits network topology-independent performances that match the centralized `SGD`.[1] Conversely, the results in the existing literature imply that `GT-DSGD` is always worse than the centralized `SGD`. When the global function further satisfies the Polyak-Łojasiewics (PL) condition, it is shown that `GT-DSGD` converges linearly up to a steady-state error with appropriate constant step-sizes. With a family of stochastic approximation step-sizes, we show that `GT-DSGD` achieves the optimal global sublinear rate with probability one and the asymptotically optimal sublinear rate in mean.

4.1 Introduction

We consider decentralized non-convex optimization where n nodes cooperate to solve the following problem:

$$\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{4.1}$$

such that each function $f_i : \mathbb{R}^p \to \mathbb{R}$ is local and private to node i and the nodes communicate over a balanced directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{1, \cdots, n\}$ is the set of node indices and $\mathcal{E}$ is the collection of ordered pairs $(i, j), i, j \in \mathcal{V}$, such that node j sends information to node i. Throughout the chapter, we assume that each local f_i is smooth and non-convex. We focus on an *online*[2] setup where data samples are collected in

[1] Reference [149] shows that centralized `SGD` is optimal for this problem class.

[2] We note that "online" sometimes refers to time-varying functions, which is different from the problem setup in this thesis.

real-time and hence each node i only has access to a noisy sample $\mathbf{g}_i$ of the true gradient at each iteration, such that $\mathbf{g}_i$ is an unbiased estimate of ∇f_i with bounded variance. Problems of this nature have found significant interest in signal processing, machine learning, and control [7, 14].

4.1.1 Related work

Based on the classical stochastic gradient descent (`SGD`) [7], a well-known solution to Problem (4.1) is decentralized `SGD` (`DSGD`) [37, 39]. However, the convergence of `DSGD` for non-convex problems has only been established under certain regularity assumptions such as uniformly bounded difference between local and global gradients [2, 41, 43], or coercivity of each local function [42]. It has also been observed that if the data distributions across the nodes are heterogeneous, the practical performance of `DSGD` degrades significantly [14, 20, 67]. One notable line of work towards improving the performance of `DSGD` is EXTRA [68] and Exact Diffusion [115], where the convergence under the stochastic non-convex setting is established without the aforementioned regularity assumptions [3]; however, they require the weight matrix to be symmetric and the smallest eigenvalue is lower bounded by $-1/3$. Another family of algorithms to eliminate the performance limitation of `DSGD` is based on gradient tracking, introduced in [55, 65], where the basic idea is to replace the local gradients with a tracker of the global gradient ∇F. Decentralized first-order methods with gradient tracking have been well studied under exact gradients, where relevant work can be found, e.g., in [52–54, 56, 129]. However, the convergence behavior of gradient tracking methods has many unanswered questions when it comes to non-convex online stochastic problems [126, 150].

4.1.2 Main contributions

This chapter considers `GT-DSGD` [67], that adds gradient tracking to `DSGD`, for online stochastic non-convex problems and rigorously develops novel results, key insights, and new analysis techniques that fill the theory gaps in the existing literature on gradient tracking methods [67, 126, 150]. The main contributions are described in the following.

- **General smooth non-convex problems.** We explicitly characterize the non-asymptotic, transient and steady-state performance of `GT-DSGD` and derive the conditions under which they are comparable to that of the centralized minibatch `SGD`. In particular, we show that its non-asymptotic mean-squared rate is network-independent and further matches the centralized minibatch `SGD` when the number of iterations is large enough. In sharp contrast, the existing results in [126, 150] suggest that the convergence rate and steady-state performance of `GT-DSGD` are always network-dependent and therefore are strictly worse than that of the centralized minibatch `SGD`; see Section 4.3.1 for details.

- **Problems satisfying the global Polyak-Łojasiewicz (PL) condition.** We analyze `GT-DSGD` when the global (smooth non-convex) function F further satisfies the PL condition. For both constant and decaying step-sizes, we explicitly characterize the non-asymptotic, transient and steady-state behaviors in expectation, and establish the conditions under which they are comparable to that of the centralized minibatch `SGD`. We further establish global sublinear convergence rates on almost every sample path. The obtained sample path-wise rates are order-optimal. To the best of our knowledge, these are the first results on path-wise convergence rate for online decentralized stochastic optimization under non-convexity, thus generalizing prior results in the decentralized stochastic approximation literature, e.g., [151], where the convergence analysis is mostly performed under assumptions of local convexity. As special cases, these results improve the current state-of-the-art on exact gradient methods under the PL condition [143] and stochastic strongly convex problems [67]; see Section 4.3.2 for details.

- **Technical analysis.** We emphasize that the analysis techniques in this work are substantially different from the existing ones [67], [150], [126] and may be applied to other gradient methods built upon similar principles. We describe a few key features in the following. We establish tighter bounds on the stochastic gradient tracking process, by exploiting the unbiasedness of the online stochastic gradients, based on which all convergence theorems are derived; see Section 4.5.1.2. To prove the convergence under general non-convexity, we characterize a descent inequality explicitly with network consensus errors and further show that the cumulative consensus errors along the algorithm path are dominated by the cumulative descent effect of the local gradients; see Section 4.5.1.3. Towards the convergence analysis under the global PL condition, we derive the uniform boundedness of gradient tracking errors that is crucial in simplifying the ensuing analysis; see Lemma 4.5.18. Subsequently, we construct an appropriate stochastic process that forms an almost supermartingale [152] to prove sublinear rates on almost every sample path; see Section 4.5.2.2. To develop the convergence results in mean under the global PL condition, we use the analytical tools developed for recursive processes with time-varying step-sizes; see Section 4.5.2.3.

The rest of the chapter is organized as follows. Section 4.2 describes the assumptions and the `GT-DSGD` algorithm. In Section 4.3, we present the main results and discuss the contributions of this chapter in the context of the current state-of-the-art, whereas Section 4.3.1 and 4.3.2 respectively focus on the general non-convex and the PL case. We present detailed numerical experiments in Section 4.4 to demonstrate the main theoretical results in this chapter. Section 4.5 presents the detailed proofs of the main theorems in this chapter. Section 4.5.1 establishes general bounds on the stochastic gradient tracking process and proves the convergence for smooth non-convex functions. Sections 4.5.2.1, 4.5.2.2, and 4.5.2.3 provide the convergence

analysis under the PL condition on top of the results obtained in Section 4.5.1. In particular, Sections 4.5.2.1 and 4.5.2.3 focus on the convergence in mean with constant and decaying step-sizes respectively while Section 4.5.2.2 focuses on the almost sure convergence. Section 4.6 concludes the chapter.

We use lowercase bold letters to denote vectors and uppercase bold letters for matrices. The matrix, $\mathbf{I}_d$ (resp. $\mathbf{O}_d$), represents the $d \times d$ identity (resp. zero matrix); $\mathbf{1}_d$ and $\mathbf{0}_d$ are the d-dimensional column vectors of all ones and zeros, respectively. We denote $[\mathbf{x}]_i$ as the i-th entry of a vector $\mathbf{x}$. The Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} \otimes \mathbf{B}$. We use $\|\cdot\|$ to denote the Euclidean norm of a vector or the spectral norm of a matrix. For a matrix $\mathbf{X}$, we use $\rho(\mathbf{X})$ to denote its spectral radius, $\mathbf{X}^*$ to denote its adjugate, $\det(\mathbf{X})$ to denote its determinant, $[\mathbf{X}]_{i,j}$ to denote its (i,j)th element and $\mathrm{diag}(\mathbf{X})$ as the diagonal matrix that consists of the diagonal entries of $\mathbf{X}$. Matrix-vector inequalities are interpreted in the entry-wise sense. We use $\sigma(\cdot)$ to denote the σ-algebra generated by the random variables and/or sets in its argument.

4.2 Assumptions and the `GT-DSGD` Algorithm

We are interested in finding a first-order stationary point of Problem (4.1) via local computation and communication at each node. We first enlist the necessary assumptions that are standard in the literature [5,7,20,67].

Assumption 4.2.1 (Objective functions). *Each f_i is L-smooth, i.e., $\exists L > 0$ s.t. $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Moreover, F is bounded below, i.e., $F^* := \inf_{\mathbf{x}} F(\mathbf{x}) > -\infty$.*

Assumption 4.2.2 (Network model). *The directed communication network is strongly-connected and admits a primitive doubly-stochastic weight matrix $\underline{\mathbf{W}} = \{\underline{w}_{ir}\} \in \mathbb{R}^{n \times n}$.*

We consider iterative processes that generate at each node i a sequence of state vectors $\{\mathbf{x}_k^i : k \geq 0\}$, where $\mathbf{x}_0^i$ is assumed to be a constant. At each iteration k, each node i is able to call the local oracle that returns a stochastic gradient $\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$, where $\boldsymbol{\xi}_k^i$ is a random vector in $\mathbb{R}^q$ and $\mathbf{g}_i : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^p$ is a Borel-measurable function. For example, $\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$ may be considered as the stochastic gradient evaluated at the state $\mathbf{x}_k^i$ with the data sample $\boldsymbol{\xi}_k^i$ observed at node i and iteration k. We work with a rich enough probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and define the natural filtration as, $\forall k \geq 1$,

$$\mathcal{F}_k := \sigma\left(\left\{\boldsymbol{\xi}_t^i : 0 \leq t \leq k-1, i \in \mathcal{V}\right\}\right), \qquad \mathcal{F}_0 := \{\Omega, \phi\},$$

where ϕ is the empty set. The intuitive meaning of $\mathcal{F}_k$ is that it contains the historical information of the algorithm iterates in question up to iteration $k-1$.

Algorithm 5 `GT-DSGD` at each node i

Require: $\mathbf{x}_0^i$; $\{\alpha_k\}$; $\{\underline{w}_{ir}\}$; $\mathbf{y}_i^0 = \mathbf{0}_p$; $\mathbf{g}_r\left(\mathbf{x}_{-1}^r, \boldsymbol{\xi}_{-1}^r\right) := \mathbf{0}_p$.

1: **for** $k = 0, 1, \ldots,$ **do**

$$\mathbf{y}_{k+1}^i = \sum_{r=1}^n \underline{w}_{ir}\left(\mathbf{y}_k^r + \mathbf{g}_r(\mathbf{x}_k^r, \boldsymbol{\xi}_k^r) - \mathbf{g}_r(\mathbf{x}_{k-1}^r, \boldsymbol{\xi}_{k-1}^r)\right)$$

$$\mathbf{x}_{k+1}^i = \sum_{r=1}^n \underline{w}_{ir}\left(\mathbf{x}_k^r - \alpha_k \mathbf{y}_{k+1}^r\right)$$

2: **end for**

Assumption 4.2.3 (**Oracle model**). *The stochastic gradient process* $\{\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i) : \forall k \geq 0, \forall i \in \mathcal{V}\}$ *satisfies:*

- $\mathbb{E}\left[\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i) | \mathcal{F}_k\right] = \nabla f_i(\mathbf{x}_k^i), \forall k \geq 0, \forall i \in \mathcal{V}$;
- $\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i) - \nabla f_i(\mathbf{x}_k^i)\right\|^2 | \mathcal{F}_k\right] \leq \nu_i^2, \forall k \geq 0, \forall i \in \mathcal{V}$, *for some constant* $\nu_i > 0$;
- *The family* $\left\{\boldsymbol{\xi}_k^i : \forall k \geq 0, \forall i \in \mathcal{V}\right\}$ *of random vectors is independent.*

We denote $\nu_a^2 := \frac{1}{n}\sum_{i=1}^n \nu_i^2$, the average of the variance of local stochastic gradients. We are also interested in the case when *the global objective function* F further satisfies the Polyak-Łojasiewicz (PL) condition that was introduced in [5].

Assumption 4.2.4. $\exists \mu > 0$ *s.t. the global* $F : \mathbb{R}^p \to \mathbb{R}$ *satisfies* $2\mu\left(F(\mathbf{x}) - F^*\right) \leq \left\|\nabla F(\mathbf{x})\right\|^2, \forall \mathbf{x} \in \mathbb{R}^p$.

When Assumption 4.2.4 holds, we denote $\kappa := \frac{L}{\mu} \geq 1$, which can be interpreted as the condition number of F; see Lemma 4.5.12. Note that under the PL condition, every stationary point $\mathbf{x}^*$ of F is a global minimum of F, while F is not necessarily convex. Assumption 4.2.4 holds, e.g., in certain reinforcement learning problems [153], see [5, 147] for more details.

`GT-DSGD`, introduced in [67] for smooth strongly convex problems and formally described in Algorithm 5, recursively descends in the direction of an auxiliary variable $\mathbf{y}_k^i$ at each node, instead of the local stochastic gradient $\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$. The auxiliary variable $\mathbf{y}_k^i$ is constructed under the dynamic average consensus principle [51] and tracks a time-varying signal $\sum_i \mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$, which mimics the global gradient; see [14, 67] for further intuition and explanation. We note that `GT-DSGD` uses the adapt-then-combine (ATC) structure [37] resulting in improved stability of the algorithm.

4.3 Main results

In this section, we present our main convergence results for `GT-DSGD` and compare them with the corresponding state-of-the-art. For analysis purposes and the ease of presentation of main results, we let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{g}_k$,

all in $\mathbb{R}^{np}$, respectively concatenate $\mathbf{x}_k^i$'s, $\mathbf{y}_k^i$'s, $\mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i)$'s, and write `GT-DSGD` in the following matrix form: $\forall k \geq 0$,

$$\mathbf{y}_{k+1} = \mathbf{W}\left(\mathbf{y}_k + \mathbf{g}_k - \mathbf{g}_{k-1}\right), \tag{4.2a}$$

$$\mathbf{x}_{k+1} = \mathbf{W}\left(\mathbf{x}_k - \alpha_k \mathbf{y}_{k+1}\right), \tag{4.2b}$$

where $\mathbf{W} = \underline{\mathbf{W}} \otimes \mathbf{I}_p$. We denote the exact averaging matrix as $\mathbf{J} := (\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes \mathbf{I}_p$ and $\lambda := \|\mathbf{W} - \mathbf{J}\|$, which characterizes the network connectivity. Under Assumption 4.2.2, we have $\lambda \in [0, 1)$; see [36]. For convenience, we let $\nabla\mathbf{f}_k \in \mathbb{R}^{np}$ concatenate all local exact gradients $\nabla f_i(\mathbf{x}_k^i)$'s and denote

$$\overline{\mathbf{x}}_k := \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{x}_k, \quad \overline{\mathbf{y}}_k := \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{y}_k,$$
$$\overline{\nabla\mathbf{f}}_k := \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\nabla\mathbf{f}_k, \quad \overline{\mathbf{g}}_k := \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{g}_k.$$

We assume without loss of generality that $\mathbf{x}_0^i = \mathbf{x}_0^r, \forall i, r \in \mathcal{V}$.

4.3.1 General smooth non-convex functions

In this subsection, we are concerned with the convergence of `GT-DSGD` for general smooth non-convex functions.

Theorem 4.3.1. *Let Assumptions 4.2.1, 4.2.2, and 4.2.3 hold and consider `GT-DSGD` under a constant step-size $\alpha_k = \alpha, \forall k \geq 0$, such that $0 < \alpha \leq \min\left\{1, \frac{1-\lambda^2}{12\lambda}, \frac{(1-\lambda^2)^2}{4\sqrt{6}\lambda^2}\right\}\frac{1}{2L}$, then, $\forall K > 1$,*

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right]}_{\text{Mean-squared stationary gap}} \leq \underbrace{\frac{4(F(\overline{\mathbf{x}}_0) - F^*)}{\alpha K} + \frac{2\alpha\nu_a^2 L}{n}}_{\text{Centralized minibatch SGD}} + \underbrace{\frac{448\alpha^2L^2\lambda^2\nu_a^2}{(1-\lambda^2)^3} + \frac{64\alpha^2L^2\lambda^4}{(1-\lambda^2)^3 K}\frac{\|\nabla\mathbf{f}_0\|^2}{n}}_{\text{Decentralized network effect}}.$$

Further, $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F(\mathbf{x}_k^i)\|^2\right]$ decays at the rate of $\mathcal{O}(\frac{1}{K})$ up to a steady-state error such that

$$\limsup_{K\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \leq \underbrace{\frac{2\alpha\nu_a^2 L}{n}}_{\text{Centralized minibatch SGD}} + \underbrace{\frac{448\alpha^2L^2\lambda^2\nu_a^2}{(1-\lambda^2)^3}}_{\text{Decentralized network effect}}.$$

Theorem 4.3.1 is proved in Section 4.5.1.

Remark 4.3.1 (Transient and steady-state performance). Theorem 4.3.1 explicitly characterizes the non-asymptotic performance of `GT-DSGD` for general smooth non-convex functions with an appropriate constant step-size. In particular, the stationary gap of `GT-DSGD` for any finite number of iterations K is bounded by the sum of four terms. The first two terms are independent of the network spectral gap $1 - \lambda$ and match the complexity of the centralized minibatch `SGD` up to constant factors [7]. The third and the fourth terms depend on $1 - \lambda$ reflecting the decentralized network and are in the order of $\mathcal{O}(\alpha^2)$. This is a much tighter

characterization compared with the existing results [126, 150] on **GT-DSGD** and leads to provably faster non-asymptotic rate, see Remark 4.3.2 below. Theorem 4.3.1 also shows that as $K \to \infty$, the stationary gap of **GT-DSGD** decays sublinearly at the rate of $\mathcal{O}(1/K)$ up to a steady-state error. It can be observed that if $\alpha = \mathcal{O}\big(\frac{(1-\lambda)^3}{\lambda^2 nL}\big)$, then the steady state stationary gap of **GT-DSGD** matches that of the centralized minibatch **SGD** up to constant factors. The existing analysis [126], however, suggests that under the same choice of the step-size α, the steady state stationary gap of **GT-DSGD** is strictly worse than the centralized minibatch **SGD**.

The following corollary of Theorem 4.3.1 is concerned with the non-asymptotic convergence rate of **GT-DSGD** over a finite time horizon for general smooth non-convex functions.

Corollary 4.3.1. *Let Assumptions 4.2.1, 4.2.2, and 4.2.3 hold and suppose that* $\|\nabla \mathbf{f}_0\|^2 = \mathcal{O}(n)$. *Setting* $\alpha = \sqrt{n/K}$ *in Theorem 4.3.1, for* $K \geq 4nL^2 \max\left\{1, \frac{144\lambda^2}{(1-\lambda^2)^2}, \frac{96\lambda^4}{(1-\lambda^2)^4}\right\}$, *we obtain:*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \leq \underbrace{\frac{4(F(\overline{\mathbf{x}}_0)-F^*)}{\sqrt{nK}}+\frac{2\nu_a^2 L}{\sqrt{nK}}}_{\textit{Centralized minibatch SGD}} + \underbrace{\frac{448n\lambda^2\nu_a^2L^2}{(1-\lambda^2)^3K}+\frac{64L^2\lambda^4\|\nabla \mathbf{f}_0\|^2}{(1-\lambda^2)^3K^2}}_{\textit{Decentralized network effect}}.$$

Thus, if K *further satisfies that* $K \geq K_{nc} := \mathcal{O}\left(\frac{n^3\lambda^4L^2}{(1-\lambda)^6}\right)$, *then we have*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] = \mathcal{O}\left(\frac{\nu_a^2 L}{\sqrt{nK}}\right).$$

Remark 4.3.2 (**Non-asymptotic mean-squared rate and transient time for network independence**). Corollary 4.3.1 shows that if the number of iterations is large enough, i.e., $K \geq K_{nc}$, by setting $\alpha = \frac{\sqrt{n}}{\sqrt{K}}$, the non-asymptotic rate of **GT-DSGD** matches that of the centralized minibatch **SGD** up to factors of universal constants. This discussion shows that, in the regime that $K \geq K_{nc}$, **GT-DSGD** achieves a network-independent linear speedup compared with the centralized minibatch **SGD** that processes all data at a single node. In other words, the number of stochastic gradient computations required to achieve an approximate stationary point is reduced by a factor of $1/n$ at each node in the network. These results significantly improve the existing convergence guarantees of **GT-DSGD** for general smooth non-convex functions [126, 150]. In particular, references [126, 150] show that if $\alpha = \frac{c_0}{\sqrt{K}}$, where K is large enough and c_0 is some positive constant, **GT-DSGD** achieves the convergence rate of $\frac{c_1}{\sqrt{K}}$, where c_1 is a function of the network spectral gap $(1-\lambda)$. The convergence results in [126, 150] thus suggest that the rate of **GT-DSGD** is always network-dependent and is strictly worse than that of the centralized minibatch **SGD** and hence fail to characterize the network-independent performance of **GT-DSGD**.

Remark 4.3.3 (**Comparison with DSGD**). We observe from Corollary 4.3.1 that the convergence of **GT-DSGD** is robust to the difference between the local and the global functions. In other words, **GT-DSGD**

outperforms **DSGD** when data distributions across the nodes are significantly heterogeneous, since the convergence rate of the latter explicitly depends on a factor that measures the heterogeneity between the local and the global functions [2]. However, the transient time for **GT-DSGD** to achieve network independent performance has a network dependence of $\mathcal{O}((1-\lambda)^{-6})$ which is worse than that of **DSGD** where the dependence is $\mathcal{O}((1-\lambda)^{-4})$. Moreover, we note that **GT-DSGD** requires two consecutive rounds of communication per node per iteration to update the state and the gradient tracker variables respectively, compared to **DSGD**.

4.3.2 Smooth non-convex functions under PL condition

In this subsection, we discuss the performance of **GT-DSGD** when the global objective function F further satisfies the PL condition. We begin with the case of constant step-size.

Theorem 4.3.2. *Let Assumption 4.2.1, 4.2.2, 4.2.3 and 4.2.4 hold. If the step-size $\alpha_k = \alpha, \forall k \geq 0$, satisfies*

$$0 < \alpha \leq \overline{\alpha} := \min\left\{\frac{1}{2L}, \frac{(1-\lambda^2)^2}{42\lambda^2 L}, \frac{1-\lambda^2}{24\lambda L\kappa^{1/4}}, \frac{1-\lambda^2}{2\mu}\right\},$$

then $\mathbb{E}[\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2]$ and $\mathbb{E}[F(\overline{\mathbf{x}}_k) - F^*]$ decay linearly at $\mathcal{O}((1-\mu\alpha)^k)$ up to a steady-state error such that

$$\limsup_{k\to\infty} \mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right] \leq \frac{288\lambda^4\alpha^5 L^3\kappa\nu_a^2}{n(1-\lambda^2)^4} + \frac{144\lambda^2\alpha^2\nu_a^2}{(1-\lambda^2)^3},$$
$$\limsup_{k\to\infty} \mathbb{E}\left[F(\overline{\mathbf{x}}_k) - F^*\right] \leq \frac{3\alpha\kappa\nu_a^2}{2n} + \frac{72\lambda^2\alpha^2\kappa L\nu_a^2}{(1-\lambda^2)^3}.$$

Moreover, $\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[F(\mathbf{x}_k^i) - F^*\right]$ decays linearly at $\mathcal{O}((1-\mu\alpha)^k)$ up to a steady-state error such that

$$\limsup_{k\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[F(\mathbf{x}_k^i) - F^*\right] = \underbrace{\mathcal{O}\left(\frac{\alpha\kappa\nu_a^2}{n}\right)}_{\text{Centralized minibatch SGD}} + \underbrace{\mathcal{O}\left(\frac{\lambda^2\alpha^2\kappa L\nu_a^2}{(1-\lambda)^3}\right)}_{\text{Decentralized network effect}}.$$

Theorem 4.3.2 is proved in Section 4.5.2.1. Here we highlight some key features in the following remarks.

Remark 4.3.4 (**Transient and steady-state performance**)**.** Theorem 4.3.2 shows that when the global objective function F satisfies the PL condition and the constant step-size α is less than $\overline{\alpha}$, the optimality gap of **GT-DSGD** decays linearly up to a steady-state error that is the sum of two terms. The first term is independent of the network and matches that of the centralized minibatch **SGD** up to constant factors, while the second term is due to the network and is controlled by $\mathcal{O}(\alpha^2)$. In contrast to [67], which requires a stronger assumption that the global objective function is strongly convex, we note that our stability range of the step-size α is larger by a factor of $\mathcal{O}(\kappa^{5/12})$; this relaxed upper bound on α further leads to a faster linear convergence when exact gradients are available, see Remark 4.3.5. Next, it can be verified from Theorem 4.3.2 that to match the steady-state error performance of the centralized minibatch **SGD** (up to constant factors), it suffices to choose the step-size α in **GT-DSGD** such that $\alpha = \mathcal{O}\big(\frac{(1-\lambda)^3}{\lambda^2 nL}\big)$, which is larger

by a factor of $\mathcal{O}(\kappa)$ than the corresponding result in [67]; in other words, Theorem 4.3.2 demonstrates a tighter and faster convergence rate to achieve the same steady-state error.

Remark 4.3.5 (**Global linear convergence under exact gradient oracle**). Theorem 4.3.2 further shows that when the exact gradient oracle is available at each node, i.e., $\nu_i^2 = 0, \forall i \in \mathcal{V}$, `GT-DSGD` reduces to its deterministic counterpart [54, 56, 65] and achieves global linear convergence to an optimal solution with an appropriate constant step-size. In other words, when $\alpha = \overline{\alpha}$, it achieves an q-accurate optimal solution in $\mathcal{O}\big(\max\big\{\kappa, \frac{\lambda^2\kappa}{(1-\lambda)^2}, \frac{\lambda\kappa^{5/4}}{1-\lambda}, \frac{1}{1-\lambda}\big\}\log\frac{1}{q}\big)$ iterations. This result improves upon the state-of-the-art gradient computation and communication complexity under the PL condition [143]. The gradient computation complexity can be further improved to $\mathcal{O}\big(\kappa\log\frac{1}{\epsilon}\big)$ by performing $\mathcal{O}\big(\frac{1}{1-\lambda}\log\frac{\kappa}{1-\lambda}\big)$ rounds of consensus communication at each iteration. This gradient computation complexity result matches the state-of-the-art [69] on decentralized exact gradient methods (without Nesterov acceleration), which further requires a stronger assumption that each local function is convex and the global function is strongly convex. In contrast, we only require the PL condition on the global objective F.

We now proceed to the case of decaying step-sizes. The next result shows the sample path-wise performance of `GT-DSGD` under a family of stochastic approximation step-sizes [154], i.e., $\alpha_k > 0$, $\sum_{k=0}^{\infty}\alpha_k = \infty$, and $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$, which enables the exact sublinear convergence in contrast to the inexact linear convergence under a constant step-size.

Theorem 4.3.3. *Let Assumptions 4.2.1, 4.2.2, 4.2.3, and 4.2.4 hold. Consider the step-size sequence $\{\alpha_k\}$ such that $\alpha_k = \delta(k+\varphi)^{-\epsilon}, \forall k \geq 0$, where $\epsilon \in (0.5, 1]$, $\delta \geq 1/\mu$, and $\varphi \geq \max\big\{(\delta/\overline{\alpha})^{1/\epsilon}, \frac{4}{1-\lambda^2}\big\}$ for $\overline{\alpha}$ given in Theorem 4.3.2. Then $\forall i, j \in \mathcal{V}$ and for arbitrarily small $\epsilon_1 > 0$, we have:*

$$\mathbb{P}\left(\sum_{k=0}^{\infty} k^{2\epsilon-1-\epsilon_1}\left\|\mathbf{x}_k^i - \mathbf{x}_k^j\right\|^2 < \infty\right) = 1,$$
$$\mathbb{P}\Big(\lim_{k\to\infty} k^{2\epsilon-1-\epsilon_1}\big(F(\mathbf{x}_k^i) - F^*\big) = 0\Big) = 1.$$

Theorem 4.3.3 is proved in Section 4.5.2.2.

Remark 4.3.6 (**Global sublinear rate on almost every sample path**). Theorem 4.3.3 guarantees that `GT-DSGD` exhibits a global sublinear convergence on almost every sample path, under decaying step-sizes, when the global function F satisfies the PL condition. This result is of significant practical value in that it is applicable to every instantiation of the algorithm while the expectation type convergence only characterizes, roughly speaking, the performance on average. Furthermore, in the case of general non-degenerate variances (see Assumption 4.2.3), these path-wise rates are order-optimal, in the sense of polynomial time decay; this follows by considering the stochastic approximation reformulation of the optimization problem (i.e.,

the problem of obtaining zeros of the gradient function $\nabla F(\mathbf{x})$) and invoking standard central limit type arguments, see [154].) To the best of our knowledge, Theorem 4.3.3 is the first to show path-wise convergence for online decentralized stochastic optimization under non-convexity, thus generalizing prior results in the decentralized stochastic approximation and optimization literature, such as [151], where such analysis is performed under assumptions of local convexity.

Finally, we consider the convergence rate of **GT-DSGD** in expectation when $\alpha_k = \mathcal{O}(1/k), \forall k \geq 0$.

Theorem 4.3.4. *Let Assumptions 4.2.1, 4.2.2, 4.2.3, and 4.2.4 hold. Consider the step-size sequence $\{\alpha_k\}$ such that $\alpha_k = \beta(k+\gamma)^{-1}, \forall k \geq 0$, where $\beta > 2/\mu$, and $\gamma \geq \max\left\{\frac{\beta}{\overline{\alpha}}, \frac{8}{1-\lambda^2}\right\}$ for $\overline{\alpha}$ given in Theorem 4.3.2. We have: $\forall k \geq 0$,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[F(\mathbf{x}_k^i)-F^*\right] \leq \underbrace{\frac{2L\nu_a^2\beta^2}{n(\mu\beta-1)(k+\gamma)}}_{\text{Centralized minibatch SGD}} + \underbrace{\frac{2\left(F(\overline{\mathbf{x}}_0)-F^*\right)}{(k/\gamma+1)^{\mu\beta}} + \frac{3L^2\widehat{x}\beta^3}{n(\mu\beta-2)(k+\gamma)^2}}_{\text{Decentralized network effect}},$$

where $\widehat{x}$ is a positive constant given in (4.63).

The non-asymptotic rate in Theorem 4.3.4 shows that **GT-DSGD** asymptotically achieves network independent $\mathcal{O}(1/k)$ rate in mean when the global objective function F satisfies the PL condition, matching the $\Omega(1/k)$ oracle lower bound [7]. The following corollary examines the number of transient iterations required to achieve network-independence under specific choices of parameter β and γ in Theorem 4.3.4.

Corollary 4.3.2. *Let Assumptions 4.2.1, 4.2.2, 4.2.3, and 4.2.4 hold. Set $\beta = 6/\mu$ and $\gamma = \max\left\{\frac{6}{\mu\overline{\alpha}}, \frac{8}{1-\lambda^2}\right\}$ in Theorem 4.3.4 and suppose that $\|\nabla \mathbf{f}_0\|^2 = \mathcal{O}(n)$. Then we have:*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[F(\mathbf{x}_k^i)-F^*\right] = \mathcal{O}\left(\frac{\kappa^2\left(F(\overline{\mathbf{x}}_0)-F^*\right)}{k^2} + \frac{\kappa\nu_a^2}{n\mu k}\right),$$

if k is large enough such that $k \gtrsim K_{PL}$, where

$$K_{PL} := \frac{\lambda^2 n\kappa}{(1-\lambda)^3} + \frac{\lambda\kappa^{5/4}}{1-\lambda} + \kappa + \frac{\lambda^{3/2}\kappa^{11/8}}{(1-\lambda)^{3/2}} + \frac{\kappa^{-1/2}}{(1-\lambda)^{3/2}} + \frac{\lambda^2 n\kappa^{1/2}L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda)^2\nu_a^2}.$$

Theorem 4.3.4 and Corollary 4.3.2 are proved in Section 4.5.2.3.

Remark 4.3.7 (Transient time for network independent rate). Corollary 4.3.2 shows after K_{PL} iterations, the convergence rate of **GT-DSGD** matches that of the centralized minibatch **SGD** [7] up to constant factors and therefore achieves an asymptotic linear speedup. We now compare this transient time with the existing literature. First, Ref. [67] shows that, under the strong convexity of F, **GT-DSGD** asymptotically converges at $\mathcal{O}(1/k)$; however, the convergence rate derived in [67] depends on arbitrary constants and therefore the transient time is not clear. Second, recent work [116] shows that when each local function f_i is

Table 4.1: A summary of the datasets used in numerical experiments, available at https://www.openml.org/.

Dataset	train	dimension	classes
a9a	48,832	124	2
w8a	60,000	301	2
creditcard	100,000	30	2
Fashion-MNIST	60,000	785	10
CIFAR-10	50,000	3073	10
STL-10	5,000	27649	10

strongly convex, the corresponding transient time of `DSGD` is $\mathcal{O}\big(n\kappa^6(1-\lambda)^{-2}\big)$. Our results on the transient time K_{PL} therefore significantly improve upon the dependence of the condition number κ under weaker assumptions on the objective functions, while being moderately worse in terms of the network dependence.

4.4 Numerical Experiments

In this section, we present numerical experiments to demonstrate the main theoretical results in Section 4.3 with the help of learning problems on real-world datasets, summarized in Table 4.1, and minimizing certain synthetic functions to illustrate the PL condition. We consider three different graph topologies, i.e., a directed exponential graph with 16 nodes, an undirected grid graph with 16 nodes, and an undirected geometric graph with 100 nodes; see Fig. 4.1. The primitive doubly stochastic weights are set to be equal for the exponential graph and are generated by the Metroplis rule [27] for the grid and the geometric graphs. The second largest singular values λ associated with the weight matrices of these graphs are $0.6, 0.93$ and 0.99, respectively. Towards the stochastic gradient oracle, we consider two different setups: (i) each node has access to a finite collection of data samples and the stochastic gradient is computed with respect to one randomly selected data sample at each iteration; (ii) each node has access to the gradient of its local function subject to random noise, with zero-mean and bounded variance, at each iteration. The performance metric of interest is the average of global function values across the nodes $\frac{1}{n}\sum_{i=1}^{n} F(\mathbf{x}_k^i)$, which we refer to as *loss*, versus the number of epochs[3] in (i) and the number of iterations in (ii). We manually optimize the parameters of all algorithms across all experiments to achieve their best performances.

To study the convergence behavior of `GT-DSGD`, we conduct three different experiments: binary classification with non-convex logistic regression [137], multiclass classification with neural networks, and minimizing synthetic non-convex functions that satisfy the global PL condition. We compare the performance of `GT-DSGD` with `DSGD` [2] to illustrate the advantages of the former in the setting of heterogeneous data distributions

[3]Each epoch is one effective pass of local data samples at each node.

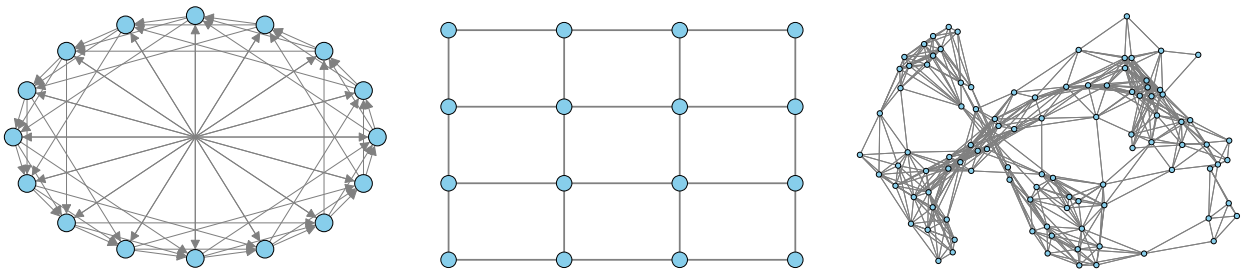

Figure 4.1: A directed exponential graph with 16 nodes, an undirected grid graph with 16 nodes, and an undirected geometric graph with 100 nodes.

across the nodes; moreover, we use the centralized minibatch **SGD** as the benchmark to illustrate the scenarios in which **GT-DSGD** achieves a network-independent performance. The experimental results are described in the next subsections. It can be verified that the numerical results of **GT-DSGD** are consistent with the theory in this chapter.

4.4.1 Non-convex logistic regression for binary classification

We first consider a binary classification problem with the help of a non-convex logistic regression model [137]. Specifically, the decentralized optimization problem of interest is given by $\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}) + r(\mathbf{x})$, such that

$$f_i(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^m \log\left[1 + e^{-(\mathbf{x}^\top \boldsymbol{\theta}_{ij})\xi_{ij}}\right],\ r(\mathbf{x}) = \sum_{d=1}^p \frac{R[\mathbf{x}]_d^2}{1+[\mathbf{x}]_d^2},$$

where $\boldsymbol{\theta}_{i,j}$ is the feature vector, $\xi_{i,j}$ is the corresponding binary label, and $r(\mathbf{x})$ is a non-convex regularizer with $R = 10^{-4}$.

We compare the performance of **GT-DSGD** over the directed exponential and the grid graphs, both with 16 nodes, to the centralized **SGD** with a minibatch size of 16. We consider the best possible constant step-size for both algorithms. The numerical results over the a9a, w8a, and creditcard datasets are shown in Fig. 4.2. It can be observed that, across all datasets, the convergence behavior of **GT-DSGD** matches that of the centralized minibatch **SGD** and is independent of the underlying graph topology, as long as the total number of iterations is large enough. This observation is consistent with Corollary 4.3.1, demonstrating the network-independent convergence of **GT-DSGD** under an appropriate constant step-size for general smooth non-convex functions.

4.4.2 Neural network for multiclass classification

We next compare the performance of **DSGD** (without gradient tracking) and **GT-DSGD**, both with a constant step-size, when the data distributions across the nodes are significantly heterogeneous. To this aim, we consider a harsh problem setup where the data samples are distributed over the 100-node geometric graph

in Fig. 4.1 such that each node has the same number of data samples and the samples belong to only one or two classes (out of 10 possible classes). We consider decentralized training of a neural network with one fully connected hidden layer of 64 neurons and sigmoid activation. The experimental results over the Fashion-MNIST, CIFAR-10, and STL-10 datasets are shown in Fig. 4.3. We observe that **GT-DSGD** significantly outperforms **DSGD** in this setting, demonstrating the robustness of **GT-DSGD** to heterogeneous data across the nodes; see also Remark 4.3.3.

4.4.3 Synthetic functions that satisfy the global PL condition

Finally, we show the performance of **GT-DSGD** when the global function satisfies the PL condition and compare it with **DSGD** and the centralized minibatch **SGD**. In particular, each local function is chosen as $f_i(x) = x^2 + 3\sin^2(x) + a_i x\cos(x)$, such that $\sum_{i=1}^n a_i = 0$ and $a_i \neq 0, \forall i \in \mathcal{V}$, leading to the global function $F(x) = x^2 + 3\sin^2(x)$, which is clearly non-convex and further satisfies the PL condition [147]. It can be verified that each local function is highly nonlinear and significantly different from the global function; see Fig. 4.4. We inject random Gaussian noise with mean 0 and the standard deviation 0.5 to the gradient computation at each node. The corresponding numerical results can be found in Fig. 4.5, where the experiments in the first three plots of Fig. 4.5 are performed over the directed exponential graph with 16 nodes. It can be observed from the first plot of Fig. 4.5 that **GT-DSGD** achieves inexact linear convergence under constant step-sizes; moreover, a smaller step-size leads to a smaller steady-state error but at a slower rate. Compared with the convergence of **DSGD** under constant step-sizes shown in the second plot of Fig. 4.5, **GT-DSGD** achieves a smaller steady-state error much faster benefiting from gradient tracking that effectively exploits the global geometry. The third plot of Fig. 4.5 shows that **GT-DSGD** achieves exact sublinear convergence to the optimal solution with decaying step-sizes of the form $\alpha_k = (k+3)^{-\tau}$ under different values of τ chosen in $(0.5, 1]$. Clearly, a larger τ leads to a faster rate as Theorem 4.3.3 suggests. Finally, we observe from the last plot of Fig. 4.5 that the convergence rate of **GT-DSGD** with $\tau = 1$ matches that of the centralized minibatch **SGD** with the same decaying step-size after a small number of transient iterations over different graphs. This phenomenon demonstrates the asymptotically network-independent and optimal $\mathcal{O}(1/k)$ rate achieved by **GT-DSGD**. This observation is consistent with Theorem 4.3.4.

4.5 Convergence analysis

4.5.1 The general non-convex case

It is straightforward to verify that the random variables generated by **GT-DSGD** are square-integrable and that $\mathbf{x}_k, \mathbf{y}_k$ are $\mathcal{F}_k$-measurable and $\mathbf{g}(\mathbf{x}_k, \boldsymbol{\xi}_k)$ is $\mathcal{F}_{k+1}$-measurable, $\forall k$. In this section, we derive general bounds

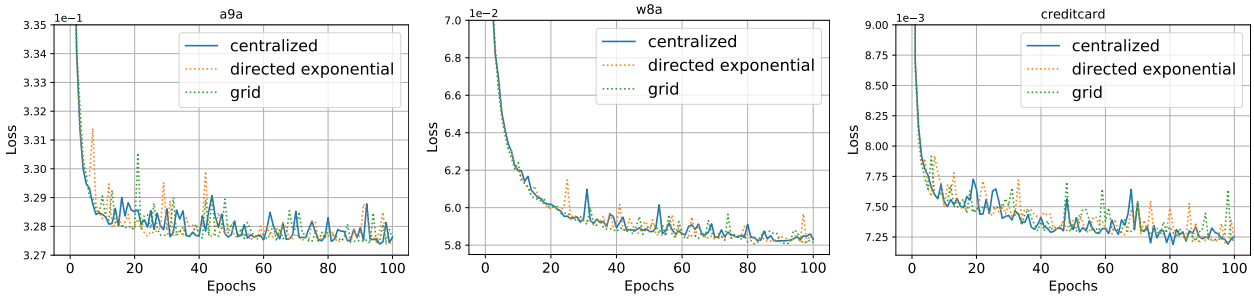

Figure 4.2: The performance of **GT-DSGD** for non-convex logistic regression over different graphs and comparison with the centralized minibatch **SGD** on the a9a, w8a and creditcard datasets.

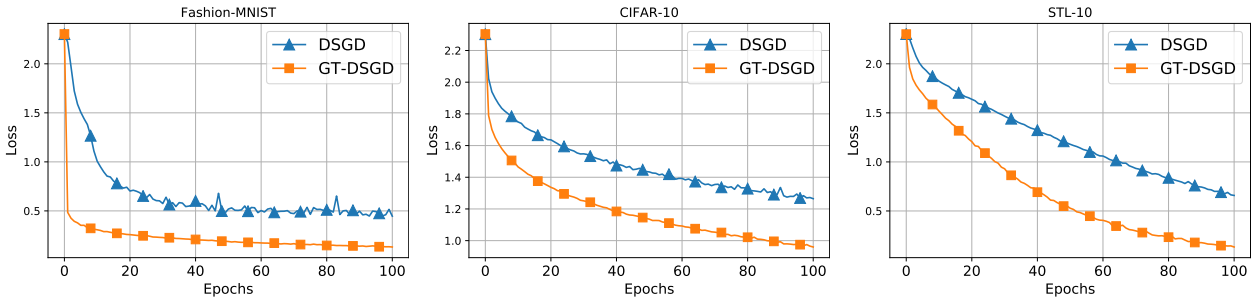

Figure 4.3: Performance comparison between **GT-DSGD** and **DSGD** for one-hidden-layer neural network under heterogeneous data distributions across the nodes on the Fashion-MNIST, CIFAR-10 and STL-10 datasets.

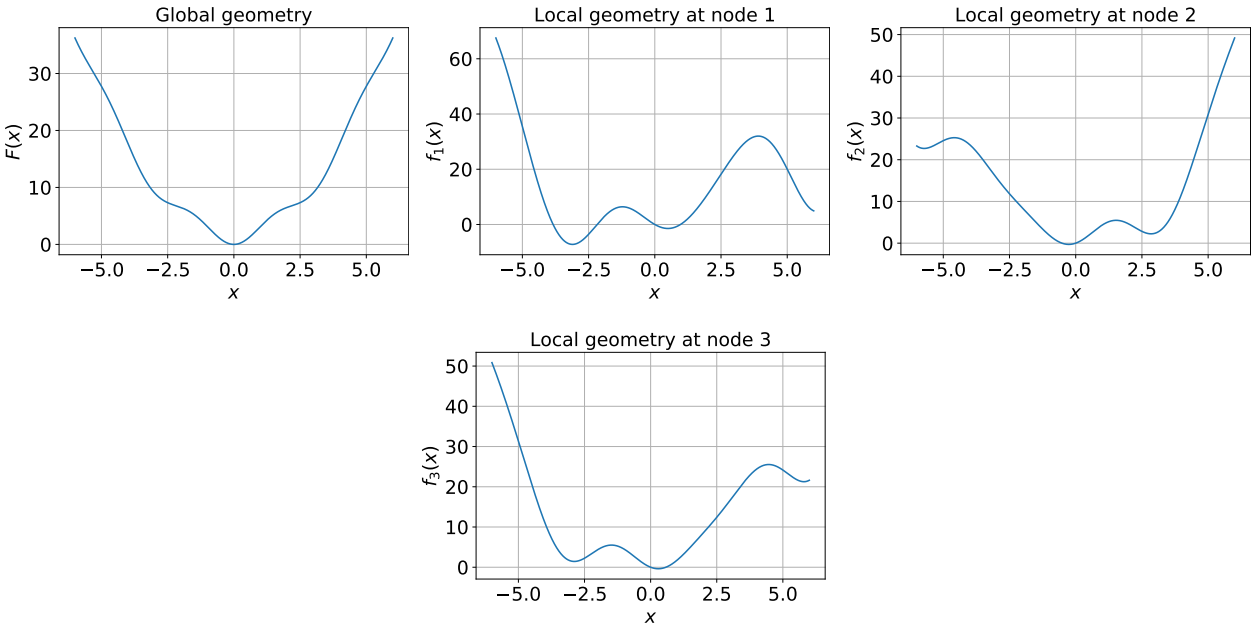

Figure 4.4: The global and local geometries in the experiment with synthetic functions that satisfy the global PL condition.

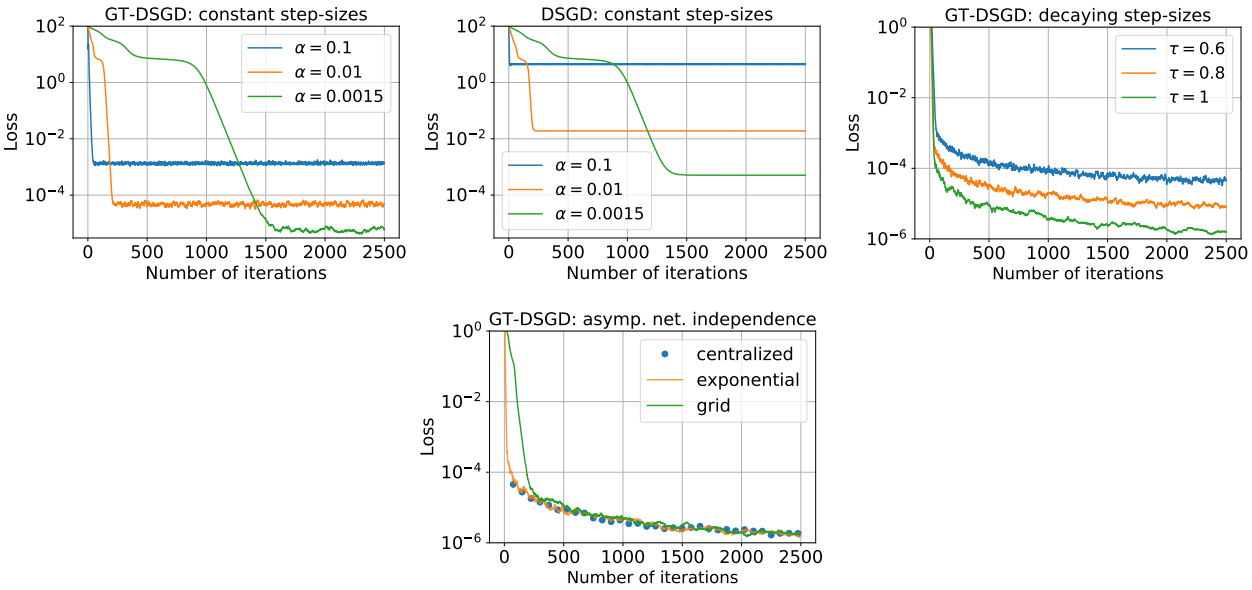

Figure 4.5: Convergence of **GT-DSGD** and **DSGD** under the global PL condition: Inexact linear convergence with different constant step-sizes α, exact sublinear convergence of **GT-DSGD** with decaying step-sizes $\alpha_k = (k+3)^{-\tau}$ under different values of τ, exact sublinear convergence of **GT-DSGD** over different graphs in comparison with the centralized minibatch **SGD** with the decaying step-size $\alpha_k = (k+3)^{-1}$.

on the stochastic gradient tracking process, which may be of independent interest, and prove Theorem 4.3.1. We start by presenting some standard results on decentralized stochastic gradient tracking algorithms; their proofs can be found, e.g., in [24, 56, 67].

Lemma 4.5.1. *Under Assumption 4.2.1-4.2.3, We have the following:*

(a) $\|\mathbf{W}\mathbf{x} - \mathbf{J}\mathbf{x}\| \leq \lambda \|\mathbf{x} - \mathbf{J}\mathbf{x}\|, \forall \mathbf{x} \in \mathbb{R}^{np}$.

(b) $\overline{\mathbf{y}}_{k+1} = \overline{\mathbf{g}}_k, \forall k \geq 0$.

(c) $\left\|\overline{\nabla \mathbf{f}}_k - \nabla F(\overline{\mathbf{x}}_k)\right\|^2 \leq \frac{L^2}{n} \|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2, \forall k \geq 0$.

(d) $\mathbb{E}[\langle \mathbf{g}_i(\mathbf{x}_k^i, \boldsymbol{\xi}_k^i) - \nabla f_i(\mathbf{x}_k^i), \mathbf{g}_r(\mathbf{x}_k^r, \boldsymbol{\xi}_k^r) - \nabla f_r(\mathbf{x}_k^r)\rangle | \mathcal{F}_k] = 0$, $\forall k \geq 0, \forall i, r \in \mathcal{V}$ *such that* $i \neq r$.

(e) $\mathbb{E}\big[\|\overline{\mathbf{g}}_k - \overline{\nabla \mathbf{f}}_k\|^2 | \mathcal{F}_k\big] \leq \nu_a^2/n, \forall k \geq 0$.

As a consequence of the state update of **GT-DSGD** described in (4.2b) and Lemma 4.5.1(b), we have: $\forall k \geq 0$,

$$\overline{\mathbf{x}}_{k+1} = \overline{\mathbf{x}}_k - \alpha_k \overline{\mathbf{y}}_{k+1} = \overline{\mathbf{x}}_k - \alpha_k \overline{\mathbf{g}}_k, \tag{4.3}$$

i.e., the mean state $\overline{\mathbf{x}}_k$ of the network proceeds in the direction of the average of local stochastic gradients $\overline{\mathbf{g}}_k$. The following lemma provides several useful relations on the consensus process of the state vectors across the network [24].

Lemma 4.5.2. *Let Assumption 4.2.2 hold. We have the following inequalities:* $\forall k \geq 0$,

$$\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\|^2 \leq \frac{1+\lambda^2}{2}\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2 + \frac{2\alpha_k^2\lambda^2}{1-\lambda^2}\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2.$$

$$\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\|^2 \leq 2\lambda^2\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2 + 2\alpha_k^2\lambda^2\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2.$$

$$\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\| \leq \lambda\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2 + \alpha_k\lambda\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|.$$

4.5.1.1 A descent inequality

In this subsection, we establish a key descent inequality that characterizes the expected decrease of the value of the global objective function F over each iteration in light of (4.3).

Lemma 4.5.3. *Let Assumptions 4.2.1-4.2.3 hold. If* $0 < \alpha_k \leq \frac{1}{2L}$, *then we have:* $\forall k \geq 0$,

$$\mathbb{E}\left[F(\overline{\mathbf{x}}_{k+1})|\mathcal{F}_k\right] \leq F(\overline{\mathbf{x}}_k) - \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k)\|^2 - \frac{\alpha_k}{4}\left\|\overline{\nabla\mathbf{f}}_k\right\|^2 + \frac{\alpha_k L^2}{2}\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n} + \frac{\alpha_k^2 L\nu_a^2}{2n}.$$

Proof. Since F is L-smooth, we have [5]: $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle\nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{4.4}$$

Setting $\mathbf{y} = \overline{\mathbf{x}}_{k+1}$ and $\mathbf{x} = \overline{\mathbf{x}}_k$ in (4.4) to obtain: $\forall k \geq 0$,

$$F(\overline{\mathbf{x}}_{k+1}) \leq F(\overline{\mathbf{x}}_k) - \alpha_k\left\langle\nabla F(\overline{\mathbf{x}}_k), \overline{\mathbf{g}}_k\right\rangle + \frac{\alpha_k^2 L}{2}\|\overline{\mathbf{g}}_k\|^2.$$

Conditioning on $\mathcal{F}_k$, by $\mathbb{E}[\overline{\mathbf{g}}_k|\mathcal{F}_k] = \overline{\nabla\mathbf{f}}_k$, obtains: $\forall k \geq 0$,

$$\begin{aligned}
\mathbb{E}[F(\overline{\mathbf{x}}_{k+1})|\mathcal{F}_k] &\leq F(\overline{\mathbf{x}}_k) - \alpha_k\left\langle\nabla F(\overline{\mathbf{x}}_k), \overline{\nabla\mathbf{f}}_k\right\rangle + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}_k\|^2|\mathcal{F}_k\right] \\
&= F(\overline{\mathbf{x}}_k) - \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k)\|^2 - \frac{\alpha_k}{2}\left\|\overline{\nabla\mathbf{f}}_k\right\|^2 + \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k) - \nabla\mathbf{f}_k\|^2 + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}_k\|^2|\mathcal{F}_k\right] \\
&\leq F(\overline{\mathbf{x}}_k) - \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k)\|^2 - \frac{\alpha_k}{2}\|\overline{\nabla\mathbf{f}}_k\|^2 + \frac{\alpha_k L^2}{2n}\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2 + \frac{\alpha_k^2 L}{2}\mathbb{E}\left[\|\overline{\mathbf{g}}_k\|^2|\mathcal{F}_k\right],
\end{aligned} \tag{4.5}$$

where the equality above uses $\langle\mathbf{x}, \mathbf{y}\rangle = \frac{1}{2}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2), \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, and the last inequality is due to Lemma 4.5.1(c). For the last term in (4.5), note that: $\forall k \geq 0$,

$$\mathbb{E}\left[\|\overline{\mathbf{g}}_k\|^2|\mathcal{F}_k\right] = \mathbb{E}\left[\left\|\overline{\mathbf{g}}_k - \overline{\nabla\mathbf{f}}_k + \overline{\nabla\mathbf{f}}_k\right\|^2|\mathcal{F}_k\right] = \mathbb{E}\left[\left\|\overline{\mathbf{g}}_k - \overline{\nabla\mathbf{f}}_k\right\|^2|\mathcal{F}_k\right] + \|\overline{\nabla\mathbf{f}}_k\|^2 \leq \nu_a^2/n + \left\|\overline{\nabla\mathbf{f}}_k\right\|^2, \tag{4.6}$$

where the second equality uses that $\overline{\nabla\mathbf{f}}_k$ is $\mathcal{F}_k$-measurable and $\mathbb{E}[\overline{\mathbf{g}}_k|\mathcal{F}_k] = \overline{\nabla\mathbf{f}}_k$, and the last inequality uses Lemma 4.5.1(e). We now use (4.6) in (4.5) to obtain: $\forall k \geq 0$,

$$\mathbb{E}[F(\overline{\mathbf{x}}_{k+1})|\mathcal{F}_k] \leq F(\overline{\mathbf{x}}_k) - \frac{\alpha_k}{2}\|\nabla F(\overline{\mathbf{x}}_k)\|^2 + \frac{\alpha_k^2 L\nu_a^2}{2n} - \frac{\alpha_k(1-\alpha_k L)}{2}\|\overline{\nabla\mathbf{f}}_k\|^2 + \frac{\alpha_k L^2}{2n}\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2.$$

The proof follows by noting that $1 - \alpha_k L \geq \frac{1}{2}$, if $0 < \alpha_k \leq \frac{1}{2L}$, $\forall k \geq 0$, in the inequality above. □

Compared with the corresponding descent inequality for the centralized stochastic gradient descent, see, e.g., [5, 7], the descent inequality for `GT-DSGD` derived in Lemma 4.5.3 has an additional network consensus error term $\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|$. We therefore seeks for means to control this perturbation in order to establish the convergence of `GT-DSGD`. We will bound the consensus and the gradient tracking error jointly.

4.5.1.2 Bounding the gradient tracking error

In this subsection, we analyze the gradient tracking process.

Lemma 4.5.4. *Let Assumption 4.2.1-4.2.3 hold. We have:* $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\,\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\|^2\,\right] \leq \lambda^2\mathbb{E}\left[\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2\,\right] + \lambda^2\mathbb{E}\left[\,\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2\,\right]\\ + 2\mathbb{E}\left[\langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_k - \mathbf{g}_k)\rangle\right]\\ + 2\mathbb{E}\left[\langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_{k+1} - \nabla\mathbf{f}_k)\rangle\right]\end{aligned}$$

Proof. Using the gradient tracking update (4.2a), and the fact that $\mathbf{W}\mathbf{J} = \mathbf{J}\mathbf{W} = \mathbf{J}$, we have: $\forall k \geq 0$,

$$\begin{aligned}&\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\|^2\\ =\;& \|\mathbf{W}\,(\mathbf{y}_{k+1} + \mathbf{g}_{k+1} - \mathbf{g}_k) - \mathbf{J}\,(\mathbf{y}_{k+1} + \mathbf{g}_{k+1} - \mathbf{g}_k)\|^2\\ =\;& \|\mathbf{W}\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1} + (\mathbf{W}-\mathbf{J})\,(\mathbf{g}_{k+1} - \mathbf{g}_k)\|^2\\ =\;& \|\mathbf{W}\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2 + \|(\mathbf{W}-\mathbf{J})\,(\mathbf{g}_{k+1} - \mathbf{g}_k)\|^2 + 2\,\langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1},\;\; (\mathbf{W}-\mathbf{J})\,(\mathbf{g}_{k+1} - \mathbf{g}_k)\rangle\\ \leq\;& \lambda^2\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2 + \lambda^2\,\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2 + 2\underbrace{\langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\mathbf{g}_{k+1} - \mathbf{g}_k)\rangle}_{C_1},\end{aligned} \tag{4.7}$$

where the last inequality is due to Lemma 4.5.1(a). Towards C_1, since $\mathbf{y}_{k+1}$ and $\mathbf{g}_k$ are $\mathcal{F}_{k+1}$-measurable, we have: $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[C_1|\mathcal{F}_{k+1}\right] &= \langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_{k+1} - \mathbf{g}_k)\rangle\\ &= \langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_k - \mathbf{g}_k)\rangle + \langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_{k+1} - \nabla\mathbf{f}_k)\rangle\,.\end{aligned} \tag{4.8}$$

The proof then follows by taking the expectation on (4.7) and using (4.8) in the resulting inequality. □

Next, we bound the terms in Lemma 4.5.4. For the second term in Lemma 4.5.4, we have the following.

Lemma 4.5.5. *Let Assumption 4.2.1-4.2.3 hold. We have:* $\forall k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\,\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2\,\right] \leq 18L^2\mathbb{E}\left[\,\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2\,\right] + 6n\alpha_k^2L^2\mathbb{E}\left[\,\|\overline{\mathbf{g}}_k\|^2\,\right]\\ + 12\alpha_k^2L^2\lambda^2\mathbb{E}\left[\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2\,\right] + 3n\nu_a^2.\end{aligned}$$

Proof. Since both $\nabla \mathbf{f}_{k+1}$ and $\mathbf{g}_k$ are $\mathcal{F}_{k+1}$-measurable and $\mathbb{E}[\mathbf{g}_{k+1}|\mathcal{F}_{k+1}] = \nabla \mathbf{f}_{k+1}$, we have: $\forall k \geq 0$,

$$\begin{aligned}
\mathbb{E}\big[\,\|\mathbf{g}_{k+1} - \mathbf{g}_k\|^2\,\big] &= \mathbb{E}\big[\,\|\mathbf{g}_{k+1} - \nabla \mathbf{f}_{k+1}\|^2\,\big] + \mathbb{E}\big[\,\|\nabla \mathbf{f}_{k+1} - \mathbf{g}_k\|^2\,\big], \\
&\leq n\nu_a^2 + \mathbb{E}\big[\,\|\nabla \mathbf{f}_{k+1} - \mathbf{g}_k\|^2\,\big] \\
&\leq n\nu_a^2 + 2\mathbb{E}\big[\,\|\nabla \mathbf{f}_{k+1} - \nabla \mathbf{f}_k\|^2\,\big] + 2\mathbb{E}\big[\,\|\nabla \mathbf{f}_k - \mathbf{g}_k\|^2\,\big] \\
&\leq 3n\nu_a^2 + 2L^2 \underbrace{\mathbb{E}\big[\,\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2\,\big]}_{C_2}
\end{aligned} \tag{4.9}$$

where the first inequality uses Assumption 4.2.3 and the last inequality uses Assumption 4.2.3 and the L-smoothness of each f_i. Towards C_2, we have: $\forall k \geq 0$,

$$\begin{aligned}
C_2 &= \mathbb{E}\big[\,\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1} + \mathbf{J}\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_k + \mathbf{J}\mathbf{x}_k - \mathbf{x}_k\|^2\,\big] \\
&\leq 3\mathbb{E}\big[\,\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\|^2\,\big] + 3n\alpha_k^2\mathbb{E}\big[\,\|\overline{\mathbf{g}}_k\|^2\,\big] + 3\mathbb{E}\big[\,\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2\,\big] \\
&\leq 9\mathbb{E}\big[\,\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2\,\big] + 3n\alpha_k^2\mathbb{E}\big[\,\|\overline{\mathbf{g}}_k\|^2\,\big] + 6\alpha_k^2\lambda^2\mathbb{E}\big[\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2\,\big],
\end{aligned} \tag{4.10}$$

where the second inequality uses (4.3) and the last inequality uses Lemma 4.5.2. The proof follows by using (4.10) in (4.9). □

For the third term in Lemma 4.5.4, we have the following.

Lemma 4.5.6. *Let Assumption 4.2.1-4.2.3 hold. We have:* $\forall k \geq 0$,

$$\mathbb{E}\left[\langle (\mathbf{W} - \mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})\,(\nabla \mathbf{f}_k - \mathbf{g}_k)\rangle\right] \leq \nu_a^2.$$

Proof. Using the fact that $\mathbf{J}(\mathbf{W} - \mathbf{J}) = \mathbf{O}_{np}$ and the gradient tracking update (4.2a), we have: $\forall k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\langle (\mathbf{W} - \mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})\,(\nabla \mathbf{f}_k - \mathbf{g}_k)\rangle\,|\mathcal{F}_k\right] \\
&= \mathbb{E}\left[\langle \mathbf{W}\mathbf{y}_{k+1}, (\mathbf{W} - \mathbf{J})\,(\nabla \mathbf{f}_k - \mathbf{g}_k)\rangle\,|\mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left\langle \mathbf{W}^2\,(\mathbf{y}_k + \mathbf{g}_k - \mathbf{g}_{k-1}), (\mathbf{W} - \mathbf{J})\,(\nabla \mathbf{f}_k - \mathbf{g}_k)\right\rangle|\mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left\langle \mathbf{W}^2\mathbf{g}_k, (\mathbf{W} - \mathbf{J})\,(\nabla \mathbf{f}_k - \mathbf{g}_k)\right\rangle|\mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left\langle \mathbf{W}^2\,(\mathbf{g}_k - \nabla \mathbf{f}_k), (\mathbf{W} - \mathbf{J})\,(\nabla \mathbf{f}_k - \mathbf{g}_k)\right\rangle|\mathcal{F}_k\right] \\
&= \mathbb{E}\left[(\mathbf{g}_k - \nabla \mathbf{f}_k)^\top(\mathbf{J} - \mathbf{W}^\top\mathbf{W}^2)\,(\mathbf{g}_k - \nabla \mathbf{f}_k)\,|\mathcal{F}_k\right],
\end{aligned} \tag{4.11}$$

where the third and the fourth equality exploit the fact that the random vectors $\mathbf{y}_k$, $\mathbf{g}_{k-1}$ and $\nabla \mathbf{f}_k$ are $\mathcal{F}_k$-

measurable and that $\mathbb{E}[\mathbf{g}_k|\mathcal{F}_k] = \nabla\mathbf{f}_k$. In light of Lemma 4.5.1(d), (4.11) reduces to

$$\begin{aligned}
&\mathbb{E}\left[\langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_k - \mathbf{g}_k)\rangle\,|\mathcal{F}_k\right] \\
&= \mathbb{E}\left[(\mathbf{g}_k - \nabla\mathbf{f}_k)^\top \mathrm{diag}(\mathbf{J} - \mathbf{W}^\top\mathbf{W}^2)\,(\mathbf{g}_k - \nabla\mathbf{f}_k)\,|\mathcal{F}_k\right] \\
&\leq \mathbb{E}\left[(\mathbf{g}_k - \nabla\mathbf{f}_k)^\top \mathrm{diag}(\mathbf{J})\,(\mathbf{g}_k - \nabla\mathbf{f}_k)\,|\mathcal{F}_k\right], \\
&= \mathbb{E}\left[\|\mathbf{g}_k - \nabla\mathbf{f}_k\|^2|\mathcal{F}_k\right]/n
\end{aligned} \tag{4.12}$$

where the inequality holds since $\mathrm{diag}(\mathbf{W}^\top\mathbf{W}^2)$ is nonnegative. The proof follows by using Assumption 4.2.3 in (4.12) and taking the expectation on the resulting inequality. □

For the last term in Lemma 4.5.4, we have the following.

Lemma 4.5.7. *Let Assumption 4.2.1-4.2.3 hold. We have:* $\forall k \geq 0$,

$$\begin{aligned}
\langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_{k+1} - \nabla\mathbf{f}_k)\rangle \leq\ & (\lambda\alpha_k L + 0.5\eta_1 + \eta_2)\lambda^2\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2 \\
& + \eta_2^{-1}\lambda^2 L^2\,\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2 + 0.5\eta_1^{-1}\lambda^2\alpha_k^2 L^2 n\,\|\overline{\mathbf{g}}_k\|^2,
\end{aligned}$$

where η_1 *and* η_2 *are arbitrary positive constants*[4].

Proof. Using $(\mathbf{W}-\mathbf{J})\mathbf{J} = \mathbf{O}_{np}$ and the Cauchy-Schwarz inequality, we have: $\forall k \geq 0$,

$$\begin{aligned}
&\langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_{k+1} - \nabla\mathbf{f}_k)\rangle \\
=\ & \langle(\mathbf{W}-\mathbf{J})\,(\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}), (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_{k+1} - \nabla\mathbf{f}_k)\rangle \\
\leq\ & \lambda^2 L\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|\,\|\mathbf{x}_{k+1} - \mathbf{x}_k\|,
\end{aligned} \tag{4.13}$$

where the last inequality uses $\|\mathbf{W}-\mathbf{J}\| = \lambda$ and the L-smoothness of each f_i. We note that, $\forall k \geq 0$,

$$\begin{aligned}
&\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \\
=\ & \|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1} + \mathbf{J}\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_k + \mathbf{J}\mathbf{x}_k - \mathbf{x}_k\| \\
\leq\ & \|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\| + \alpha_k\sqrt{n}\,\|\overline{\mathbf{g}}_k\| + \|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\| \\
\leq\ & 2\,\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\| + \alpha_k\sqrt{n}\,\|\overline{\mathbf{g}}_k\| + \alpha_k\lambda\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|.
\end{aligned} \tag{4.14}$$

where the last inequality uses Lemma 4.5.2. We use (4.14) in (4.13) to obtain: $\forall k \geq 0$,

$$\begin{aligned}
\langle(\mathbf{W}-\mathbf{J})\,\mathbf{y}_{k+1}, (\mathbf{W}-\mathbf{J})\,(\nabla\mathbf{f}_{k+1} - \nabla\mathbf{f}_k)\rangle \leq\ & \lambda^3\alpha_k L\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2 + \underbrace{(\lambda\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|)\,(\lambda\alpha_k L\sqrt{n}\,\|\overline{\mathbf{g}}_k\|)}_{C_3} \\
& + \underbrace{2(\lambda\,\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|)(\lambda L\,\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|)}_{C_4}.
\end{aligned} \tag{4.15}$$

[4]We note that η_1 and η_2 will be fixed later.

By Young's inequality, we have that

$$C_3 \leq 0.5\eta_1\lambda^2 \left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2 + 0.5\eta_1^{-1}\lambda^2\alpha_k^2L^2n\left\|\overline{\mathbf{g}}_k\right\|^2,$$

where $\eta_1 > 0$ is arbitrary, and that,

$$C_4 \leq \eta_2\lambda^2 \left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2 + \eta_2^{-1}\lambda^2L^2\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2,$$

where $\eta_2 > 0$ is arbitrary. The proof follows by Using the bounds on C_3 and C_4 in (4.15). □

With the help of auxiliary Lemmas 4.5.5-4.5.7, we now prove an upper bound on the gradient tracking error.

Lemma 4.5.8. *Let Assumption 4.2.1-4.2.3 hold. If $0 < \alpha_k \leq \frac{1-\lambda^2}{24\lambda L}$, then we have: $\forall k \geq 0$,*

$$\mathbb{E}\left[\frac{\left\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\right\|^2}{nL^2}\right] \leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\frac{\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2}{nL^2}\right] + \frac{24\lambda^2}{1-\lambda^2}\mathbb{E}\left[\frac{\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2}{n}\right] + \frac{6\lambda^2\alpha_k^2}{1-\lambda^2}\mathbb{E}\big[\left\|\overline{\nabla\mathbf{f}}_k\right\|^2\big] + \frac{6\nu_a^2}{L^2}.$$

Proof. We apply Lemma 4.5.5, 4.5.6 and 4.5.7 to Lemma 4.5.4 to obtain: $\forall k \geq 0, \forall\eta_1 > 0, \forall\eta_2 > 0$,

$$\mathbb{E}\big[\left\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\right\|^2\big] \leq \lambda^2(1 + 12\lambda^2\alpha_k^2L^2 + 2\lambda\alpha_kL + \eta_1 + 2\eta_2)\mathbb{E}\big[\left\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\right\|^2\big] + (3\lambda^2n + 2)\nu_a^2 + \left(18 + 2\eta_2^{-1}\right)\lambda^2L^2\mathbb{E}\big[\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2\big] + \left(6 + \eta_1^{-1}\right)\lambda^2\alpha_k^2L^2n\mathbb{E}\big[\left\|\overline{\mathbf{g}}_k\right\|^2\big]. \tag{4.16}$$

We set $\eta_1 = \frac{1-\lambda^2}{6\lambda^2}$ and $\eta_2 = \frac{1-\lambda^2}{12\lambda^2}$ in (4.16). It is straightforward to verify that if $0 < \alpha_k \leq \frac{1-\lambda^2}{24\lambda^2L}, \forall k \geq 0$, then we have:

$$\lambda^2(1 + 12\lambda^2\alpha_k^2L^2 + 2\lambda\alpha_kL + \eta_1 + 2\eta_2) \leq \frac{1+\lambda^2}{2}. \tag{4.17}$$

Moreover, recall from (4.6) that

$$\mathbb{E}\big[\|\overline{\mathbf{g}}_k\|^2\big] \leq \mathbb{E}\big[\|\overline{\nabla\mathbf{f}}_k\|^2\big] + \nu_a^2/n. \tag{4.18}$$

Using (4.17), (4.18), $\eta_1 = \frac{1-\lambda^2}{6\lambda^2}$ and $\eta_2 = \frac{1-\lambda^2}{12\lambda^2}$ in (4.16), we have: if $0 < \alpha_k \leq \frac{1-\lambda^2}{24\lambda^2L}$, then

$$\mathbb{E}\big[\|\mathbf{y}_{k+2} - \mathbf{J}\mathbf{y}_{k+2}\|^2\big] \leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2\big] + \left(\frac{6\lambda^2\alpha_k^2L^2}{1-\lambda^2} + 5n\right)\nu_a^2 + \frac{24\lambda^2L^2}{1-\lambda^2}\mathbb{E}\big[\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2\big] + \frac{6\lambda^2\alpha_k^2L^2n}{1-\lambda^2}\mathbb{E}\big[\|\overline{\nabla\mathbf{f}}_k\|^2\big].$$

The proof follows by $\frac{6\lambda^2\alpha_k^2L^2}{1-\lambda^2} \leq 1$ if $0 < \alpha_k \leq \frac{1-\lambda^2}{24\lambda L}, \forall k$. □

4.5.1.3 LTI dynamics

In this subsection, we establish the convergence rate of `GT-DSGD` for general smooth non-convex functions under an appropriate constant step-size such that $\alpha_k = \alpha, \forall k \geq 0$. To this end, we now jointly write Lemma 4.5.2 and 4.5.8 in the following linear-time-invariant system that characterizes the convergence of consensus and gradient tracking process.

Proposition 4.5.1. *Let Assumption 4.2.1-4.2.3 hold. If* $0 < \alpha \leq \frac{1-\lambda^2}{24\lambda L}$, *then we have the following (entry-wise) matrix-vector inequality hold:* $\forall k \geq 0$,

$$\mathbf{u}_{k+1} \leq \mathbf{G}\mathbf{u}_k + \mathbf{b}_k, \tag{4.19}$$

where the state vector $\mathbf{u}_k \in \mathbb{R}^2$, *the system matrix* $\mathbf{G} \in \mathbb{R}^{2\times 2}$ *and the input vector* $\mathbf{b}_k \in \mathbb{R}^2$ *are given by*

$$\mathbf{u}_k = \begin{bmatrix} \mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right] \\ \mathbb{E}\left[\frac{\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2}{nL^2}\right] \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \frac{1+\lambda^2}{2} & \frac{2\alpha^2\lambda^2L^2}{1-\lambda^2} \\ \frac{24\lambda^2}{1-\lambda^2} & \frac{1+\lambda^2}{2} \end{bmatrix}, \quad \mathbf{b}_k = \begin{bmatrix} 0 \\ \frac{6\lambda^2\alpha^2}{1-\lambda^2}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + \frac{6\nu_a^2}{L^2} \end{bmatrix}.$$

In light of Proposition 4.5.1, we solve the range of α such that $\rho(\mathbf{G}) < 1$, using the next lemma from [36].

Lemma 4.5.9. *Let* $\mathbf{X} \in \mathbb{R}^{d\times d}$ *be a non-negative matrix and* $\mathbf{x} \in \mathbb{R}^d$ *be a positive vector. If* $\mathbf{X}\mathbf{x} < \mathbf{x}$, *then* $\rho(\mathbf{X}) < 1$. *Moreover, if* $\mathbf{X}\mathbf{x} \leq z\mathbf{x}$, *for some* $z > 0$, *then* $\rho(\mathbf{X}) \leq z$.

Lemma 4.5.10. *If* $0 < \alpha \leq \min\left\{\frac{1-\lambda^2}{24\lambda}, \frac{(1-\lambda^2)^2}{15\lambda^2}\right\}\frac{1}{L}$, *then* $\rho(\mathbf{G}) < 1$ *and hence* $\sum_{k=0}^{\infty}\mathbf{G}^k = (\mathbf{I}_2 - \mathbf{G})^{-1}$.

Proof. In the light of Lemma 4.5.9, we solve the range of α and a positive vector $\mathbf{s} = [s_1, s_2]^\top$ such that $\mathbf{G}\mathbf{s} < \mathbf{s}$, which is equivalent to the following two inequalities:

$$\begin{cases} \frac{1+\lambda^2}{2}s_1 + \frac{2\alpha^2\lambda^2L^2}{1-\lambda^2}s_2 < s_1 \\ \frac{24\lambda^2}{1-\lambda^2}s_1 + \frac{1+\lambda^2}{2}s_2 < s_2 \end{cases} \iff \begin{cases} \alpha^2 < \frac{(1-\lambda^2)^2}{4\lambda^2L^2}\frac{s_1}{s_2} \\ \frac{s_1}{s_2} < \frac{(1-\lambda^2)^2}{48\lambda^2} \end{cases}$$

We set $s_1/s_2 = (1-\lambda^2)^2/(50\lambda^2)$ and the proof follows by using it to solve for the range of α such that the first inequality above holds. □

Now, we prove an upper bound on the accumulated consensus errors along the algorithm path as follows.

Lemma 4.5.11. *Let Assumption 4.2.1-4.2.3 hold. If* $0 < \alpha \leq \min\left\{\frac{1-\lambda^2}{24\lambda}, \frac{(1-\lambda^2)^2}{8\sqrt{6}\lambda^2}\right\}\frac{1}{L}$, *then we have:*

$$\sum_{k=0}^{K}\mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right] \leq \frac{96\alpha^4\lambda^4L^2}{(1-\lambda^2)^4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + \frac{16\alpha^2\lambda^4}{(1-\lambda^2)^3}\frac{\|\nabla \mathbf{f}_0\|^2}{n} + \frac{112\alpha^2\lambda^2\nu_a^2K}{(1-\lambda^2)^3}.$$

Proof. We recursively apply (4.19) to obtain: $\forall k \geq 1$,

$$\mathbf{u}_k \leq \mathbf{G}^k\mathbf{u}_0 + \sum_{t=0}^{k-1}\mathbf{G}^t\mathbf{b}_{k-1-t}. \tag{4.20}$$

Summing up (4.20) over k from 1 to K, we obtain: $\forall K \geq 1$,

$$\begin{aligned} \sum_{k=0}^{K}\mathbf{u}_k &\leq \sum_{k=0}^{K}\mathbf{G}^k\mathbf{u}_0 + \sum_{k=1}^{K}\sum_{t=0}^{k-1}\mathbf{G}^t\mathbf{b}_{k-1-t} \\ &\leq \left(\sum_{k=0}^{\infty}\mathbf{G}^k\right)\mathbf{u}_0 + \left(\sum_{k=0}^{\infty}\mathbf{G}^k\right)\sum_{k=0}^{K-1}\mathbf{b}_k \\ &= (\mathbf{I}_2 - \mathbf{G})^{-1}\mathbf{u}_0 + (\mathbf{I}_2 - \mathbf{G})^{-1}\sum_{k=0}^{K-1}\mathbf{b}_k. \end{aligned} \tag{4.21}$$

In light of (4.21), we next compute an (entry-wise) upper bound on $(\mathbf{I}_2 - \mathbf{G})^{-1}$ as follows. We note that if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{8\sqrt{6}\lambda^2 L}$,

$$\det(\mathbf{I}_2 - \mathbf{G}) = \frac{(1-\lambda^2)^2}{4} - \frac{48\alpha^2\lambda^4 L^2}{(1-\lambda^2)^2} \geq \frac{(1-\lambda^2)^2}{8}.$$

Using the lower bound on $\det(\mathbf{I}_2 - \mathbf{G})$ above, we have that

$$(\mathbf{I}_2 - \mathbf{G})^{-1} = \frac{(\mathbf{I}_2 - \mathbf{G})^*}{\det(\mathbf{I}_2 - \mathbf{G})} \leq \begin{bmatrix} \frac{4}{1-\lambda^2} & \frac{16\alpha^2\lambda^2 L^2}{(1-\lambda^2)^3} \\ \frac{192\lambda^2}{(1-\lambda^2)^3} & \frac{4}{1-\lambda^2} \end{bmatrix}. \tag{4.22}$$

We use (4.22) in (4.21) with $\|\mathbf{x}_0 - \mathbf{J}\mathbf{x}_0\| = 0$ to obtain: $\forall K \geq 1$,

$$\sum_{k=0}^{K} \mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right] \leq \frac{16\alpha^2\lambda^2}{(1-\lambda^2)^3}\mathbb{E}\left[\frac{\|\mathbf{y}_1 - \mathbf{J}\mathbf{y}_1\|^2}{n}\right] + \frac{96\alpha^4\lambda^4 L^2}{(1-\lambda^2)^4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla \mathbf{f}}_k\right\|^2\right] + \frac{96\alpha^2\lambda^2\nu_a^2 K}{(1-\lambda^2)^3}. \tag{4.23}$$

Finally, we use the gradient tracking update (4.2a) to obtain:

$$\begin{aligned}
&\mathbb{E}\big[\,\|\mathbf{y}_1 - \mathbf{J}\mathbf{y}_1\|^2\,\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\,\|(\mathbf{W}-\mathbf{J})\mathbf{g}_0\|^2\,\big]|\mathcal{F}_0\big] \\
&= \mathbb{E}\big[\,\|(\mathbf{W}-\mathbf{J})(\mathbf{g}_0 - \nabla\mathbf{f}_0)\|^2\,\big] + \mathbb{E}\big[\,\|(\mathbf{W}-\mathbf{J})\nabla\mathbf{f}_0\|^2\,\big] \\
&\leq \lambda^2 n\nu_a^2 + \lambda^2\,\|\nabla\mathbf{f}_0\|^2\,,
\end{aligned} \tag{4.24}$$

where the second equality uses $\mathbb{E}[\mathbf{g}_0|\mathcal{F}_0] = \nabla\mathbf{f}_0$ and that $\nabla\mathbf{f}_0$ is constant and the last inequality uses $\|\mathbf{W} - \mathbf{J}\| = \lambda$. The proof follows by using (4.24) in (4.23). □

Lemma 4.5.11 states that the accumulated consensus error may be bounded by the accumulated average of local exact gradients and the accumulated variance of stochastic gradients. We next show that this bound leads to the convergence of `GT-DSGD` for general smooth non-convex functions, i.e., Theorem 4.3.1.

Proof of Theorem 4.3.1. We take the expectation of the descent inequality in Lemma 4.5.3 and sum up the resulting inequality over k from 0 to $K-1$, $\forall K \geq 1$, to obtain: if $0 < \alpha \leq \frac{1}{2L}$,

$$\begin{aligned}
\mathbb{E}\left[F(\overline{\mathbf{x}}_K)\right] \leq \mathbb{E}\left[F(\overline{\mathbf{x}}_0)\right] - \frac{\alpha}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2\right] - \frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla\mathbf{f}}_k\right\|^2\right] + \frac{\alpha^2\nu_a^2 LK}{2n} \\
+ \frac{\alpha L^2}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right].
\end{aligned} \tag{4.25}$$

Rearranging (4.25) and using that F is bounded below by F^* obtains: if $0 < \alpha \leq \frac{1}{2L}$, $\forall K \geq 1$,

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2\right] \leq \frac{2(F(\overline{\mathbf{x}}_0) - F^*)}{\alpha} + \frac{\alpha\nu_a^2 LK}{n} - \frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla\mathbf{f}}_k\right\|^2\right] + L^2\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right]. \tag{4.26}$$

Moreover, we observe: $\forall K \geq 1$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\left(\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)-\nabla F(\overline{\mathbf{x}}_k)\right\|^2+\left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2\right]\right) \\
&\leq 2L^2\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\left\|\mathbf{x}_k-\mathbf{J}\mathbf{x}_k\right\|^2}{n}\right]+2\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}_k)\right\|^2\right],
\end{aligned}$$

where the last inequality uses the L-smoothness of F. Using (4.26) in the inequality above obtains: $\forall K \geq 1$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \leq \frac{4(F(\overline{\mathbf{x}}_0)-F^*)}{\alpha}+\frac{2\alpha\nu_a^2 LK}{n} \\
-\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla\mathbf{f}}_k\right\|^2\right]+4L^2\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\left\|\mathbf{x}_k-\mathbf{J}\mathbf{x}_k\right\|^2}{n}\right].
\end{aligned} \tag{4.27}$$

We finally apply the upper bound derived in Lemma 4.5.11 on the term of (4.27) to obtain: If $0<\alpha\leq \min\left\{\frac{1}{2},\frac{1-\lambda^2}{24\lambda},\frac{(1-\lambda^2)^2}{8\sqrt{6}\lambda^2}\right\}\frac{1}{L}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_k^i)\right\|^2\right] \\
&\leq \frac{4(F(\overline{\mathbf{x}}_0)-F^*)}{\alpha}+\frac{2\alpha\nu_a^2 LK}{n}+\frac{448\alpha^2L^2\lambda^2\nu_a^2K}{(1-\lambda^2)^3}-\left(1-\frac{384\alpha^4L^4\lambda^4}{(1-\lambda^2)^4}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\overline{\nabla\mathbf{f}}_k\right\|^2\right]+\frac{64\alpha^2L^2\lambda^4}{(1-\lambda^2)^3}\frac{\left\|\nabla\mathbf{f}_0\right\|^2}{n}.
\end{aligned}$$

Clearly, if $0<\alpha\leq\frac{1-\lambda^2}{5L\lambda}$, then $1-\frac{384\alpha^4L^4\lambda^4}{(1-\lambda^2)^4}\geq 0$, and the proof follows by dropping the negative term. □

4.5.2 The PL case

4.5.2.1 Linear convergence up to steady state error with constant step-sizes

In this section, we, built on top of the results established in Section 4.5.1, develop general bounds on the iterates of `GT-DSGD` when the global function F further satisfies the PL condition and prove Theorem 4.3.2. The following is a useful inequality that may be found in [5].

Lemma 4.5.12. *Let Assumption 4.2.1 hold. We have:* $\forall \mathbf{x}\in\mathbb{R}^p$.

$$\left\|\nabla F(\mathbf{x})\right\|^2 \leq 2L\left(F(\mathbf{x})-F^*\right).$$

Proof. By (4.4) and the fact that F is bounded below by F^*, we have $F^*\leq F\left(\mathbf{x}-L^{-1}\nabla F(\mathbf{x})\right)\leq F(\mathbf{x})-\frac{1}{2L}\left\|\nabla F(\mathbf{x})\right\|^2$, which yields the desired inequality. □

We conclude from Lemma 4.5.12 that, under Assumption 4.2.1 and 4.2.4, $\mu\leq L$ and recall $\kappa:=\frac{L}{\mu}\geq 1$. The following lemma is helpful in establishing the performance of `GT-DSGD` at each node.

Lemma 4.5.13. *Let Assumption 4.2.1 hold. We have*

$$\frac{1}{n}\sum_{i=1}^{n}(F(\mathbf{x}_k^i)-F^*)\leq 2\left(F(\overline{\mathbf{x}}_k)-F^*\right)+L\frac{\|\mathbf{x}_k-\mathbf{J}\mathbf{x}_k\|^2}{n}.$$

Proof. Setting $\mathbf{y}=\mathbf{x}_k^i$ and $\mathbf{x}=\overline{\mathbf{x}}_k$ in (4.4), we obtain

$$\begin{aligned}
F(\mathbf{x}_k^i)-F^* &\leq F(\overline{\mathbf{x}}_k)-F^*+\left\langle\nabla F(\overline{\mathbf{x}}_k),\mathbf{x}_k^i-\overline{\mathbf{x}}_k\right\rangle+\tfrac{1}{2}L\left\|\mathbf{x}_k^i-\overline{\mathbf{x}}_k\right\|^2,\\
&\leq F(\overline{\mathbf{x}}_k)-F^*+\|\nabla F(\overline{\mathbf{x}}_k)\|\left\|\mathbf{x}_k^i-\overline{\mathbf{x}}_k\right\|+\tfrac{1}{2}L\left\|\mathbf{x}_k^i-\overline{\mathbf{x}}_k\right\|^2,\\
&\leq F(\overline{\mathbf{x}}_k)-F^*+\tfrac{1}{2}L^{-1}\|\nabla F(\overline{\mathbf{x}}_k)\|^2+L\left\|\mathbf{x}_k^i-\overline{\mathbf{x}}_k\right\|^2\\
&\leq 2\left(F(\overline{\mathbf{x}}_k)-F^*\right)+L\left\|\mathbf{x}_k^i-\overline{\mathbf{x}}_k\right\|^2, \qquad (4.28)
\end{aligned}$$

where the third inequality uses Young's inequality and the last inequality is due to Lemma 4.5.12. Averaging (4.28) over i from 1 to n proves the lemma. □

We next refine several results developed in Section 4.5.1. We first use the PL inequality to in Lemma 4.5.3.

Lemma 4.5.14. *Let Assumptions 4.2.1-4.2.4 hold. If $0<\alpha_k\leq\frac{1}{2L}$, then we have: $\forall k\geq 0$,*

$$\mathbb{E}\left[\frac{F(\overline{\mathbf{x}}_{k+1})-F^*}{L}\middle|\mathcal{F}_k\right]\leq(1-\mu\alpha_k)\frac{F(\overline{\mathbf{x}}_k)-F^*}{L}+\frac{\alpha_k L}{2}\frac{\|\mathbf{x}_k-\mathbf{J}\mathbf{x}_k\|^2}{n}+\frac{\alpha_k^2\nu_a^2}{2n}.$$

Proof. The proof follows by using the PL condition in the descent inequality in Lemma 4.5.3 and then substracting F^* from both sides of the resulting inequality. □

We next use Lemma 4.5.12 to refine Lemma 4.5.8 as follows.

Lemma 4.5.15. *Let Assumption 4.2.1-4.2.3 hold. If $0<\alpha_k\leq\min\left\{\frac{1-\lambda^2}{12\lambda},1\right\}\frac{1}{2L}$, then we have: $\forall k\geq 0$,*

$$\begin{aligned}
\mathbb{E}\left[\frac{\|\mathbf{y}_{k+2}-\mathbf{J}\mathbf{y}_{k+2}\|^2}{nL^2}\right]\leq\frac{1+\lambda^2}{2}\mathbb{E}\left[\frac{\|\mathbf{y}_{k+1}-\mathbf{J}\mathbf{y}_{k+1}\|^2}{nL^2}\right]+\frac{24\lambda^2\alpha_k^2L^2}{1-\lambda^2}\mathbb{E}\left[\frac{F(\overline{\mathbf{x}}_k)-F^*}{L}\right]\\
+\frac{27\lambda^2}{1-\lambda^2}\mathbb{E}\left[\frac{\|\mathbf{x}_k-\mathbf{J}\mathbf{x}_k\|^2}{n}\right]+\frac{6\nu_a^2}{L^2}.
\end{aligned}$$

Proof. By Lemma 4.5.1(c) and Lemma 4.5.12, we have: $\forall k\geq 0$,

$$\begin{aligned}
\left\|\overline{\nabla\mathbf{f}}_k\right\|^2 &\leq 2\|\nabla F(\overline{\mathbf{x}}_k)\|^2+2\left\|\nabla F(\overline{\mathbf{x}}_k)-\overline{\nabla\mathbf{f}}_k\right\|^2\\
&\leq 4L\left(F(\overline{\mathbf{x}}_k)-F^*\right)+2L^2n^{-1}\|\mathbf{x}_k-\mathbf{J}\mathbf{x}_k\|^2. \qquad (4.29)
\end{aligned}$$

Using the inequality above in Lemma 4.5.8 to obtain: $\forall k\geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\frac{\|\mathbf{y}_{k+2}-\mathbf{J}\mathbf{y}_{k+2}\|^2}{nL^2}\right]\leq\left(\frac{24\lambda^2}{1-\lambda^2}+\frac{12\lambda^2\alpha_k^2L^2}{1-\lambda^2}\right)\mathbb{E}\left[\frac{\|\mathbf{x}_k-\mathbf{J}\mathbf{x}_k\|^2}{n}\right]+\frac{6\nu_a^2}{L^2}\\
+\frac{24\lambda^2\alpha_k^2L}{1-\lambda^2}\mathbb{E}\left[F(\overline{\mathbf{x}}_k)-F^*\right]+\frac{1+\lambda^2}{2}\mathbb{E}\left[\frac{\|\mathbf{y}_{k+1}-\mathbf{J}\mathbf{y}_{k+1}\|^2}{nL^2}\right].
\end{aligned}$$

The proof follows by $\frac{12\lambda^2\alpha_k^2L^2}{1-\lambda^2}\leq\frac{3\lambda^2}{1-\lambda^2}$ if $0<\alpha_k\leq\frac{1}{2L}$. □

We now write the inequalities in Lemma 4.5.2, 4.5.14 and 4.5.15 jointly in a linear dynamics as follows.

Proposition 4.5.2. *Let Assumption 4.2.1-4.2.4 hold. If* $0 < \alpha_k \leq \min\left\{1, \frac{1-\lambda^2}{12\lambda}, \frac{(1-\lambda^2)^2}{4\sqrt{6}\lambda^2}\right\}\frac{1}{2L}$, *then we have the following (entry-wise) matrix-vector inequality:* $\forall k \geq 0$,

$$\mathbf{v}_{k+1} \leq \mathbf{H}_k \mathbf{v}_k + \mathbf{u}_k, \tag{4.30}$$

where the state vector $\mathbf{v}_k \in \mathbb{R}^3$, *the system matrix* $\mathbf{H} \in \mathbb{R}^{3\times 3}$ *and the input vector* $\mathbf{u}_k \in \mathbb{R}^3$ *are given by*

$$\mathbf{v}_k = \begin{bmatrix} \mathbb{E}\left[\frac{\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2}{n}\right] \\ \mathbb{E}\left[\frac{F(\overline{\mathbf{x}}_k) - F^*}{L}\right] \\ \mathbb{E}\left[\frac{\|\mathbf{y}_{k+1} - \mathbf{J}\mathbf{y}_{k+1}\|^2}{nL^2}\right] \end{bmatrix}, \quad \mathbf{u}_k = \begin{bmatrix} 0 \\ \frac{\alpha_k^2\nu_a^2}{2n} \\ \frac{6\nu_a^2}{L^2} \end{bmatrix}, \quad \mathbf{H}_k = \begin{bmatrix} \frac{1+\lambda^2}{2} & 0 & \frac{2\alpha_k^2\lambda^2L^2}{1-\lambda^2} \\ \frac{\alpha_k L}{2} & 1-\mu\alpha_k & 0 \\ \frac{27\lambda^2}{1-\lambda^2} & \frac{24\lambda^2\alpha_k^2L^2}{1-\lambda^2} & \frac{1+\lambda^2}{2} \end{bmatrix}.$$

In the following lemma, we find the range of the step-size α_k such that $\rho(\mathbf{H}_k) < 1, \forall k \geq 0$, with the help of Lemma 4.5.9.

Lemma 4.5.16. *Let Assumption 4.2.1-4.2.4 hold. If the step-size sequence* α_k *satisfies for all* k *that*

$$0 < \alpha_k \leq \overline{\alpha} := \min\left\{\frac{1}{2L}, \frac{(1-\lambda^2)^2}{42\lambda^2 L}, \frac{1-\lambda^2}{24\lambda L\kappa^{1/4}}, \frac{1-\lambda^2}{2\mu}\right\}, \tag{4.31}$$

then we have: $\rho(\mathbf{H}_k) \leq 1 - \frac{\mu\alpha_k}{2} < 1$, $\forall k \geq 0$.

Proof. In the light of Lemma 4.5.9, we solve for the range of the step-size α_k and a positive vector $\boldsymbol{\delta} = [\delta_1, \delta_2, \delta_3]$ such that $\mathbf{H}_k\boldsymbol{\delta} \leq \left(1 - \frac{\mu\alpha_k}{2}\right)\boldsymbol{\delta}$, which may be written as

$$\frac{\mu\alpha_k}{2} + \frac{2\alpha_k^2\lambda^2L^2}{1-\lambda^2}\frac{\delta_3}{\delta_1} \leq \frac{1-\lambda^2}{2}, \tag{4.32}$$

$$\kappa\delta_1 \leq \delta_2, \tag{4.33}$$

$$\frac{\mu\alpha_k}{2} \leq \frac{1-\lambda^2}{2} - \frac{27\lambda^2}{1-\lambda^2}\frac{\delta_1}{\delta_3} - \frac{24\lambda^2\alpha_k^2L^2}{1-\lambda^2}\frac{\delta_2}{\delta_3}. \tag{4.34}$$

According to (4.33), we fix $\delta_1 = 1$ and $\delta_2 = \kappa$. We now impose that $0 < \alpha_k \leq \frac{1-\lambda^2}{2\mu}, \forall k \geq 0$. Then, according to (4.34), we choose $\delta_3 > 0$ such that $\frac{27\lambda^2}{1-\lambda^2}\frac{1}{\delta_3} + \frac{24\lambda^2\alpha_k^2L^2}{1-\lambda^2}\frac{\kappa}{\delta_3} \leq \frac{1-\lambda^2}{4}$. It suffices to fix $\delta_3 = \frac{108\lambda^2}{(1-\lambda^2)^2} + \frac{96\lambda^2\alpha_k^2L^2\kappa}{(1-\lambda^2)^2}$. Now, we use the fixed values of $\delta_1, \delta_2, \delta_3$ and the requirement that $0 < \alpha_k \leq \frac{1-\lambda^2}{2\mu}$ to solve the range of α_k such that (4.32) holds, i.e.,

$$\frac{216\alpha_k^2\lambda^4L^2}{(1-\lambda^2)^3} + \frac{192\alpha_k^4\lambda^4L^4\kappa}{(1-\lambda^2)^3} \leq \frac{1-\lambda^2}{4}.$$

It therefore suffices to choose α_k such that

$$0 < \alpha_k \leq \min\left\{\frac{1-\lambda^2}{6\lambda L\kappa^{1/4}}, \frac{(1-\lambda^2)^2}{42\lambda^2 L}\right\}.$$

Summarizing the obtained upper bounds on α_k in the discussion completes the proof. □

We note that $\overline{\alpha}$ defined in (4.31) is the same as the one given in Theorem 4.3.2. The following lemma drives upper bounds on several important quantities.

Lemma 4.5.17. *Let Assumption 4.2.1-4.2.4 hold. If $0 < \alpha_k \leq \overline{\alpha}$, where $\overline{\alpha}$ is given in (4.31), then we have: $\forall k \geq 0$,*

$$\left[\left(\mathbf{I}_3 - \mathbf{H}_k \right)^{-1} \mathbf{u}_k \right]_1 \leq \frac{288\lambda^4 \alpha_k^5 L^3 \kappa \nu_a^2}{n(1-\lambda^2)^4} + \frac{144\alpha_k^2 \lambda^2 \nu_a^2}{(1-\lambda^2)^3},$$
$$\left[\left(\mathbf{I}_3 - \mathbf{H}_k \right)^{-1} \mathbf{u}_k \right]_2 \leq \frac{3\alpha_k \nu_a^2}{2\mu n} + \frac{72\lambda^2 \alpha_k^2 \kappa \nu_a^2}{(1-\lambda^2)^3}.$$

Proof. By the definition of $\mathbf{H}_k$ in Proposition 4.5.2, we first compute the determinant of $(\mathbf{I}_3 - \mathbf{H}_k)$: $\forall k \geq 0$,

$$\begin{aligned} \det(\mathbf{I}_3 - \mathbf{H}_k) &= \frac{\mu\alpha_k(1-\lambda^2)^2}{4} - \frac{24\alpha_k^5 L^5 \lambda^4}{(1-\lambda^2)^2} - \frac{54\mu\alpha_k^3 L^2 \lambda^4}{(1-\lambda^2)^2} \\ &\geq \frac{\mu\alpha_k(1-\lambda^2)^2}{12}. \end{aligned}$$

if $0 < \alpha_k \leq \overline{\alpha}$, where $\overline{\alpha}$ is given in (4.31). Moreover, the adjugate of $\mathbf{I}_3 - \mathcal{H}_k$, denoted as $\underline{\mathbf{H}}^*$, is given by

$$[\underline{\mathbf{H}}^*]_{1,2} = \frac{48\lambda^4 \alpha_k^4 L^4}{(1-\lambda^2)^2}, \qquad [\underline{\mathbf{H}}^*]_{1,3} = \frac{2\mu\alpha_k^3 \lambda^2 L^2}{1-\lambda^2},$$
$$[\underline{\mathbf{H}}^*]_{2,2} \leq \frac{(1-\lambda^2)^2}{4}, \qquad [\underline{\mathbf{H}}^*]_{2,3} = \frac{\alpha_k^3 L^3 \lambda^2}{1-\lambda^2}.$$

The proof follows by $(\mathbf{I}_3 - \mathbf{H}_k)^{-1} = \underline{\mathbf{H}}^* / \det(\mathbf{I}_3 - \mathbf{H}_k)$ and the definition of $\mathbf{u}_k$ given in Proposition 4.5.2. □

We are now ready to prove Theorem 4.3.2 that characterizes the performance of `GT-DSGD` under a constant step-size.

Proof of Theorem 4.3.2. We consider a constant step-size such that $\alpha_k = \alpha, \forall k \geq 0$, with $0 < \alpha \leq \overline{\alpha}$ where $\overline{\alpha}$ is given in (4.31). We denote $\mathbf{H}_k := \mathbf{H}$ and $\mathbf{u}_k := \mathbf{u}, \forall k \geq 0$, and recursively apply (4.30) from k to 1 to obtain: $\forall k \geq 1$,

$$\mathbf{v}_k \leq \mathbf{H}^k \mathbf{v}_0 + \sum_{t=0}^{k-1} \mathbf{H}^t \mathbf{u} \leq \mathbf{H}^k \mathbf{v}_0 + \left(\mathbf{I}_3 - \mathbf{H} \right)^{-1} \mathbf{u}. \tag{4.35}$$

It is then clear that the first two statements in Theorem 4.3.2 follow by using Lemma 4.5.16 and 4.5.17 in (4.35) and the third statement in Theorem 4.3.2 follows by Lemma 4.5.13. □

4.5.2.2 Almost sure sublinear rate with stochastic approximation step-sizes

In this section, we prove Theorem 4.3.3, i.e., the almost sure sublinear convergence rates of `GT-DSGD` when the global function satisfies the PL condition under a family of stochastic approximation step-sizes. We first establish a key fact that under appropriate step-sizes, the stochastic gradient tracking errors are uniformly bounded in mean squared across all iterations. This fact will also be used in Section 4.5.2.3.

Lemma 4.5.18. *Let Assumptions 4.2.1-4.2.4 hold. If $0 < \alpha_k \leq \overline{\alpha}$, for $\overline{\alpha}$ given in (4.31), then we have:*

$$\sup_{k\geq 0} \mathbb{E}\big[\, \|\mathbf{y}_k - \mathbf{J}\mathbf{y}_k\|^2 \,\big] \leq \widehat{y},$$

where $\widehat{y}$ is a positive constant given by

$$\widehat{y} := \frac{30\lambda^2\overline{\alpha}^3 L^3\kappa\nu_a^2}{(1-\lambda^2)^2} + \frac{60n\lambda^2\overline{\alpha}^2 L^3(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda^2)^2} + \frac{16n\nu_a^2}{1-\lambda^2} + \lambda^2\|\nabla\mathbf{f}_0\|^2. \tag{4.36}$$

Proof. We prove by mathematical induction that for the state vector $\mathbf{v}_k$ defined in Proposition 4.5.2, there exists some positive constant vector $\widehat{\mathbf{v}} = [\widehat{v}_1, \widehat{v}_2, \widehat{v}_3]^\top$ such that

$$\mathbf{v}_k \leq \widehat{\mathbf{v}}, \qquad \forall k \geq 0. \tag{4.37}$$

if $0 < \alpha_k \leq \overline{\alpha}$, where $\overline{\alpha}$ is given in (4.31). We first note that in order to make (4.37) hold when $k = 0$, according to the definition of $\mathbf{v}_0$ and (4.24), it suffices to choose $\widehat{\mathbf{v}}$ such that

$$\widehat{\mathbf{v}}^\top \geq \left[0, \frac{F(\overline{\mathbf{x}}_0)-F^*}{L}, \frac{\lambda^2\nu_a^2}{L^2} + \frac{\lambda^2\|\nabla\mathbf{f}_0\|^2}{nL^2}\right]. \tag{4.38}$$

Next, we show that if $\mathbf{v}_k \leq \widehat{\mathbf{v}}$ for some $k \geq 0$ and then we also have $\mathbf{v}_{k+1} \leq \widehat{\mathbf{v}}$ with an appropriate choice of $\widehat{\mathbf{v}}$. In light of Proposition 4.5.2, we have $\mathbf{v}_{k+1} \leq \mathbf{H}_k\mathbf{v}_k + \mathbf{u}_k \leq \mathbf{H}_k\widehat{\mathbf{v}} + \mathbf{u}_k$, and hence it suffices to choose $\widehat{\mathbf{v}}$ such that $\mathbf{H}_k\widehat{\mathbf{v}} + \mathbf{u}_k \leq \widehat{\mathbf{v}}, \forall k$, which is equivalent to the following set of inequalities:

$$\frac{2\alpha_k^2\lambda^2L^2}{1-\lambda^2}\widehat{v}_3 \leq \frac{1-\lambda^2}{2}\widehat{v}_1, \tag{4.39}$$

$$\frac{\kappa}{2}\widehat{v}_1 + \frac{\alpha_k\nu_a^2}{2\mu n} \leq \widehat{v}_2, \tag{4.40}$$

$$\frac{27\lambda^2}{1-\lambda^2}\widehat{v}_1 + \frac{24\lambda^2\alpha_k^2L^2}{1-\lambda^2}\widehat{v}_2 + \frac{6\nu_a^2}{L^2} \leq \frac{1-\lambda^2}{2}\widehat{v}_3, \tag{4.41}$$

where $0 < \alpha_k \leq \overline{\alpha}$ and $\kappa = L/\mu$. First, we note that to make (4.39) hold, it suffices to choose $\widehat{v}_1$ as

$$\widehat{v}_1 = \frac{4\overline{\alpha}^2\lambda^2L^2}{(1-\lambda^2)^2}\widehat{v}_3. \tag{4.42}$$

Second, based on (4.38), (4.40), and (4.42), we choose $\widehat{v}_2$ as

$$\widehat{v}_2 = \frac{2\overline{\alpha}^2\lambda^2L^2\kappa}{(1-\lambda^2)^2}\widehat{v}_3 + \frac{\overline{\alpha}\nu_a^2}{2\mu n} + \frac{F(\overline{\mathbf{x}}_0)-F^*}{L}. \tag{4.43}$$

Third, to make (4.41) hold, it suffices to choose $\widehat{v}_3$ such that

$$\widehat{v}_3 \geq \frac{54\lambda^2}{(1-\lambda^2)^2}\widehat{v}_1 + \frac{48\lambda^2\overline{\alpha}^2L^2}{(1-\lambda^2)^2}\widehat{v}_2 + \frac{12\nu_a^2}{L^2(1-\lambda^2)}, \tag{4.44}$$

which, using (4.42) and (4.43), is equivalent to

$$\widehat{v}_3 \geq \frac{216\overline{\alpha}^2\lambda^4L^2}{(1-\lambda^2)^4}\widehat{v}_3 + \frac{96\lambda^4\overline{\alpha}^4L^4\kappa}{(1-\lambda^2)^4}\widehat{v}_3 + \frac{24\lambda^2\overline{\alpha}^3L\kappa\nu_a^2}{n(1-\lambda^2)^2} + \frac{48\lambda^2\overline{\alpha}^2L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda^2)^2} + \frac{12\nu_a^2}{L^2(1-\lambda^2)}. \tag{4.45}$$

By the definition of $\overline{\alpha}$ in (4.31), we have $\frac{216\overline{\alpha}^2\lambda^4L^2}{(1-\lambda^2)^4} \leq \frac{6}{49}$ and that $\frac{96\overline{\alpha}^4\lambda^4L^4\kappa}{(1-\lambda^2)^4} \leq \frac{1}{3456}$; therefore, to make (4.45) hold, it suffices to choose $\widehat{v}_3$ such that

$$\widehat{v}_3 \geq \frac{30\lambda^2\overline{\alpha}^3L\kappa\nu_a^2}{n(1-\lambda^2)^2} + \frac{60\lambda^2\overline{\alpha}^2L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda^2)^2} + \frac{15\nu_a^2}{L^2(1-\lambda^2)}.$$

Based on the above inequality and (4.38), we choose $\widehat{v}_3$ as

$$\widehat{v}_3 = \frac{30\lambda^2\overline{\alpha}^3L\kappa\nu_a^2}{n(1-\lambda^2)^2} + \frac{60\lambda^2\overline{\alpha}^2L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda^2)^2} + \frac{16\nu_a^2}{L^2(1-\lambda^2)} + \frac{\lambda^2\|\nabla\mathbf{f}_0\|^2}{nL^2}.$$

The induction is complete and the proof then follows by the definition of $\mathbf{v}_k$ in Proposition 4.5.2. □

We prove Theorem 4.3.3 using the Robbins-Siegmund almost supermartingale convergence theorem [152], presented as follows.

Lemma 4.5.19 (Robbins-Siegmund). *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}, \mathbb{P})$ be a filtered space. Suppose that Z_k, B_k, C_k and D_k are nonnegative and $\mathcal{F}_k$-measurable random variables such that*

$$\mathbb{E}\left[Z_{k+1}|\mathcal{F}_k\right] \leq (1+B_k)\,Z_k + C_k - D_k,\ \forall k \geq 0.$$

Then on the event $\{\sum_{k=0}^{\infty} B_k < \infty,\ \sum_{k=0}^{\infty} C_k < \infty\}$, we have that $\lim_{k\to\infty} Z_k$ exists and is finite almost surely, and that $\sum_{k=0}^{\infty} D_k < \infty$ almost surely.

We are now ready to present the proof of Theorem 4.3.3, where we construct appropriate almost supermartingales that characterize the sample path-wise convergence rate of **GT-DSGD** under a family of stochastic approximation step-sizes.

Proof of Theorem 4.3.3. We consider the step-size sequence $\{\alpha_k\}$ of the following form: $\forall k \geq 0$,

$$\alpha_k = \delta(k+\varphi)^{-\epsilon}, \quad \text{where } \delta \geq 1/\mu \text{ and } \epsilon \in (0.5, 1], \tag{4.46}$$

such that $\varphi \geq \max\{(\delta/\overline{\alpha})^{1/\epsilon}, \frac{4}{1-\lambda^2}\}$. Hence, $0 < \alpha_k \leq \overline{\alpha}$ for $\overline{\alpha}$ given in (4.31). We construct $\mathcal{F}_k$-adapted processes: $\forall k \geq 0$,

$$R_k := (k+\varphi)^{\tau}\widetilde{x}_k := (k+\varphi)^{\tau}n^{-1}\left\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\right\|^2,$$
$$Q_k := (k+\varphi)^{\tau}\Delta_k := (k+\varphi)^{\tau}L^{-1}(F(\overline{\mathbf{x}}_k) - F^*),$$

where $\tau = 2\epsilon - 1 - \epsilon_1$, where $\epsilon_1 \in (0, 2\epsilon-1)$ is an arbitrarily small constant. By $1 + x \leq e^x, \forall x \in \mathbb{R}$, we have $(k+\varphi+1)^{\tau} = (k+\varphi)^{\tau}\big(1+\frac{1}{k+\varphi}\big)^{\tau} \leq (k+\varphi)^{\tau}e^{\frac{\tau}{k+\varphi}}$. Since $0 < \frac{\tau}{k+\varphi} \leq 1$, we have: $\forall k \geq 0$,

$$(k+\varphi+1)^{\tau} \leq e(k+\varphi)^{\tau}. \tag{4.47}$$

Further, by $e^x \leq 1+x+x^2$ for $0 \leq x \leq 1$,[5] we have: $\forall k \geq 0$,

$$(k+\varphi+1)^\tau \leq \left(1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}\right)(k+\varphi)^\tau. \tag{4.48}$$

Recursion of R_k. We use Lemma 4.5.18 in Lemma 4.5.2 with the definition of α_k in (4.46) to obtain: $\forall k \geq 0$,

$$\mathbb{E}\left[\widetilde{x}_{k+1}\right] \leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\widetilde{x}_k\right] + \frac{2\lambda^2\widehat{y}}{n(1-\lambda^2)}\frac{\delta^2}{(k+\varphi)^{2\epsilon}}, \tag{4.49}$$

where $\widehat{y}$ is given in (4.36). We multiply (4.49) by $(k+\varphi+1)^\tau$ and apply (4.47) and (4.48) to obtain: $\forall k \geq 0$,

$$\mathbb{E}\left[R_{k+1}\right] \leq \underbrace{\frac{1+\lambda^2}{2}\left(1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}\right)}_{T_k}\mathbb{E}\left[R_k\right] + \frac{2e\lambda^2\widehat{y}}{n(1-\lambda^2)}\frac{\delta^2}{(k+\varphi)^{2\epsilon-\tau}}. \tag{4.50}$$

Since $\varphi \geq \frac{4}{1-\lambda^2}$, i.e., $\frac{\tau}{k+\varphi} \leq \frac{1-\lambda^2}{4}, \forall k \geq 0$, we have

$$\begin{aligned} T_k &= \left(1-\frac{1-\lambda^2}{2}\right)\left(1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}\right) \\ &\leq 1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}-\frac{1-\lambda^2}{2} \\ &\leq 1+\frac{\tau^2}{(k+\varphi)^2}-\frac{1-\lambda^2}{4}. \end{aligned} \tag{4.51}$$

Using (4.51) in (4.50), we have: $\forall k \geq 0$,

$$\mathbb{E}\left[R_{k+1}\right] \leq \left(1+\frac{\tau^2}{(k+\varphi)^2}\right)\mathbb{E}\left[R_k\right] - \frac{1-\lambda^2}{4}\mathbb{E}\left[R_k\right] + \frac{2e\lambda^2\widehat{y}}{n(1-\lambda^2)}\frac{\delta^2}{(k+\varphi)^{2\epsilon-\tau}}. \tag{4.52}$$

Note that $\sum_{k=0}^\infty (k+\varphi)^{-2} < \infty$ and $\sum_{k=0}^\infty (k+\varphi)^{\tau-2\epsilon} < \infty$ since $2\epsilon-\tau > 1$. Applying a special case of Lemma 4.5.19 for deterministic recursions in (4.52) leads to $\sum_{k=0}^\infty \mathbb{E}\left[R_k\right] < \infty$. Since R_k is nonnegative, by monotone convergence theorem, we have $\mathbb{E}\left[\sum_{k=0}^\infty R_k\right] = \sum_{k=0}^\infty \mathbb{E}\left[R_k\right] < \infty$ which implies

$$\mathbb{P}\left(\sum_{k=0}^\infty R_k < \infty\right) = 1. \tag{4.53}$$

The first statement in Theorem 4.3.3 then follows by (4.53).

Recursion of Q_k. We recall from Lemma 4.5.14: $\forall k \geq 0$,

$$\mathbb{E}\left[\Delta_{k+1}|\mathcal{F}_k\right] \leq \left(1-\frac{\mu\delta}{(k+\varphi)^\epsilon}\right)\Delta_k + \frac{L\delta}{2(k+\varphi)^\epsilon}\widetilde{x}_k + \frac{\nu_a^2}{2n}\frac{\delta^2}{(k+\varphi)^{2\epsilon}}. \tag{4.54}$$

We multiply (4.54) by $(k+\varphi+1)^\tau$ and then use (4.47) and (4.48) to obtain: $\forall k \geq 0$,

$$\mathbb{E}[Q_{k+1}|\mathcal{F}_k] \leq \underbrace{\left(1-\frac{\mu\delta}{(k+\varphi)^\epsilon}\right)\left(1+\frac{\tau}{k+\varphi}+\frac{\tau^2}{(k+\varphi)^2}\right)}_{P_k} Q_k + \frac{eL\delta}{2(k+\varphi)^\epsilon}R_k + \frac{e\nu_a^2}{2n}\frac{\delta^2}{(k+\varphi)^{2\epsilon-\tau}}. \tag{4.55}$$

[5] Note that $e^x = 1+x+x^2\sum_{k=2}^\infty \frac{x^{k-2}}{k!}, \forall x \in \mathbb{R}$. If $0 \leq x \leq 1$, then we have $e^x \leq 1+x+x^2\sum_{k=2}^\infty \frac{1}{k!} = 1+x+(e-2)x^2 \leq 1+x+x^2$.

We observe that

$$\begin{aligned} P_k &\leq 1 + \frac{\tau}{k+\varphi} + \frac{\tau^2}{(k+\varphi)^2} - \frac{\mu\delta}{(k+\varphi)^\epsilon} \\ &\leq 1 + \frac{\tau^2}{(k+\varphi)^2} - \frac{\mu\delta - \tau}{(k+\varphi)^\epsilon}. \end{aligned} \tag{4.56}$$

We use (4.56) in (4.55) to obtain: $\forall k \geq 0$,

$$\mathbb{E}\left[Q_{k+1}|\mathcal{F}_k\right] \leq \left(1 + \frac{\tau^2}{(k+\varphi)^2}\right) Q_k - \frac{\mu\delta - \tau}{(k+\varphi)^\epsilon} Q_k + \frac{eL\delta}{2(k+\varphi)^\epsilon} R_k + \frac{e\nu_a^2}{2n} \frac{\delta^2}{(k+\varphi)^{2\epsilon-\tau}}. \tag{4.57}$$

Recall that $\sum_{k=0}^{\infty}(k+\varphi)^{-2} < \infty$ and $\sum_{k=0}^{\infty}(k+\varphi)^{\tau-2\epsilon} < \infty$ since $2\epsilon - \tau > 1$. Note that $\delta \geq 1/\mu$, i.e., $\mu\delta > \tau$, applying Lemma 4.5.19 in (4.57) with the help of (4.53) gives:

$$\mathbb{P}\Big(\lim_{k\to\infty} Q_k = Q\Big) = 1, \tag{4.58}$$

where Q is some almost surely finite random variable, and

$$\mathbb{P}\left(\sum_{k=0}^{\infty} \frac{\mu\delta - \tau}{(k+\varphi)^\epsilon} Q_k < \infty\right) = 1. \tag{4.59}$$

Since $\sum_{k=0}^{\infty} \frac{\mu\delta-\tau}{(k+\varphi)^\epsilon} = \infty$, where $\epsilon \in (0.5, 1]$, we have

$$\left\{\sum_{k=0}^{\infty} \frac{\mu\delta - \tau}{(k+\varphi)^\epsilon} Q_k < \infty\right\} \subseteq \left\{\liminf_{k\to\infty} Q_k = 0\right\}, \tag{4.60}$$

where "$\subseteq$" denotes the inclusion relation for two events. By the monotonicity of $\mathbb{P}(\cdot)$, we note that (4.59) and (4.60) lead to

$$\mathbb{P}\Big(\liminf_{k\to\infty} Q_k = 0\Big) = 1. \tag{4.61}$$

From (4.61) and (4.58), we conclude that $\mathbb{P}(Q = 0) = 1$ and the proof follows by (4.53) and Lemma 4.5.13.

□

4.5.2.3 Asymptotically optimal rate in mean with $O(1/k)$ step-size

In this section, we prove Theorem 4.3.4 and Corollary 4.3.2, i.e., the asymptotically optimal convergence rate of `GT-DSGD` in expectation and the corresponding transient time to achieve network-independent performance, when the global function F satisfies the PL condition. Recall that in this context we focus on the following step-size sequence [7]:

$$\alpha_k = \frac{\beta}{k+\gamma}, \qquad \forall k \geq 0, \tag{4.62}$$

where $\beta > 0$ and $\gamma > 0$ are parameters to be restricted later. We require $\gamma \geq \beta/\overline{\alpha}$ so that $0 < \alpha_k \leq \overline{\alpha}$ for $\overline{\alpha}$ in (4.31). We first prove a non-asymptotic rate on the consensus errors.

Lemma 4.5.20. *Let Assumption 4.2.1-4.2.4 hold. If* $\gamma \geq \max\left\{\frac{\beta}{\overline{\alpha}}, \frac{8}{1-\lambda^2}\right\}$ *for* $\overline{\alpha}$ *given in* (4.31)*, then we have:* $\forall k \geq 0$*,*

$$\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2\right] \leq \frac{\widehat{x}\beta^2}{(k+\gamma)^2}. \tag{4.63}$$

where $\widehat{x} := 8\lambda^2\widehat{y}(1-\lambda^2)^{-2}$ *for* $\widehat{y}$ *given in* (4.36)*.*

Proof. We prove by induction that there exists a constant $\widehat{x}$ such that (4.63) holds. First, since $\mathbf{x}_0^i = \mathbf{x}_0^r, \forall i, r \in \mathcal{V}$, (4.63) holds trivially when $k = 0$. We next show that if (4.63) holds for some $k \geq 0$ and then it also holds for $k+1$. From Lemma 4.5.2 and 4.5.18, we have: $\forall k \geq 0$,

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{J}\mathbf{x}_{k+1}\|^2] \leq \frac{1+\lambda^2}{2}\mathbb{E}[\|\mathbf{x}_k - \mathbf{J}\mathbf{x}_k\|^2] + \frac{2\lambda^2\widehat{y}\alpha_k^2}{1-\lambda^2}.$$

Therefore, it suffices to choose $\widehat{x}$ such that $\forall k \geq 0$,

$$\frac{1+\lambda^2}{2}\frac{\widehat{x}\beta^2}{(k+\gamma)^2} + \frac{2\lambda^2\widehat{y}}{1-\lambda^2}\frac{\beta^2}{(k+\gamma)^2} \leq \frac{\widehat{x}\beta^2}{(k+\gamma+1)^2},$$

which is equivalent to

$$\frac{2\lambda^2\widehat{y}}{1-\lambda^2} \leq \left(\frac{(k+\gamma)^2}{(k+\gamma+1)^2} - \frac{1+\lambda^2}{2}\right)\widehat{x}. \tag{4.64}$$

Since the RHS of (4.64) monotonically increases with k, we suffice to choose $\widehat{x}$ such that (4.64) holds when $k = 0$, i.e.,

$$\frac{2\lambda^2\widehat{y}}{1-\lambda^2} \leq \left(\frac{\gamma^2}{(\gamma+1)^2} - \frac{1+\lambda^2}{2}\right)\widehat{x} = \left(\frac{1-\lambda^2}{2} - \frac{2\gamma+1}{(\gamma+1)^2}\right)\widehat{x}.$$

Since $\frac{2\gamma+1}{(\gamma+1)^2} \leq \frac{2}{\gamma}$, it suffices to choose $\widehat{x}$ such that $\frac{2\lambda^2\widehat{y}}{1-\lambda^2} \leq \left(\frac{1-\lambda^2}{2} - \frac{2}{\gamma}\right)\widehat{x}$. Finally, if $\gamma \geq \frac{8}{1-\lambda^2}$, it can be observed that the induction is complete by setting $\widehat{x} := 8\lambda^2\widehat{y}(1-\lambda^2)^{-2}$. □

We next present a useful lemma adapted from [116, 151, 154].

Lemma 4.5.21. *Consider the step-size sequence* $\{\alpha_k\}$ *in* (4.62)*. We have: for any nonnegative integers* a, b *such that* $0 \leq a \leq b$*,*

$$\prod_{s=a}^{b}(1-\mu\alpha_s) \leq \frac{(a+\gamma)^{\mu\beta}}{(b+\gamma+1)^{\mu\beta}}.$$

Proof. By (4.62) and $1 + x \leq e^x, \forall x \in \mathbb{R}$, we have: $0 \leq a \leq b$,

$$\prod_{s=a}^{b}(1-\mu\alpha_s) = \prod_{s=a}^{b}\left(1 - \frac{\mu\beta}{s+\gamma}\right) \leq \exp\left\{-\sum_{s=a}^{b}\frac{\mu\beta}{s+\gamma}\right\}. \tag{4.65}$$

Since $\frac{1}{s+\gamma} \geq \int_{s+\gamma}^{s+\gamma+1}\frac{1}{x}dx, \forall s \geq 0$, we have: $0 \leq a \leq b$,

$$\sum_{s=a}^{b}\frac{1}{s+\gamma} \geq \sum_{s=a}^{b}\int_{s+\gamma}^{s+\gamma+1}\frac{1}{x}dx = \log\left(\frac{b+\gamma+1}{a+\gamma}\right). \tag{4.66}$$

Applying (4.66) to (4.65) completes the proof. □

Now we are ready to prove Theorem 4.3.4 through a non-asymptotic analysis inspired by [67,93,116,151].

Proof of Theorem 4.3.4. We denote $\Psi_k := \mathbb{E}[L^{-1}(F(\overline{\mathbf{x}}_k)-F^*)]$. Using Lemma 4.5.20 in Lemma 4.5.14 gives: if $\gamma \geq \max\left\{\frac{\beta}{\alpha}, \frac{8}{1-\lambda^2}\right\}$,

$$\Psi_{k+1} \leq (1-\mu\alpha_k)\Psi_k + \widehat{u}\alpha_k^2 + \widehat{z}\alpha_k^3, \quad \forall k \geq 0, \tag{4.67}$$

where $\widehat{u}$ and $\widehat{z}$ are defined as, for $\widehat{x}$ given in (4.63),

$$\widehat{u} := \frac{\nu_a^2}{2n} \qquad \text{and} \qquad \widehat{z} := \frac{L\widehat{x}}{2n}. \tag{4.68}$$

We recursively apply (4.67) from k to 0 to obtain[6]: $\forall k \geq 1$,

$$\begin{aligned}
\Psi_k &\leq \Psi_0 \prod_{t=0}^{k-1}(1-\mu\alpha_t) + \sum_{t=0}^{k-1}\left(\left(\widehat{u}\alpha_t^2 + \widehat{z}\alpha_t^3\right)\prod_{l=t+1}^{k-1}(1-\mu\alpha_l)\right) \\
&\leq \Psi_0 \frac{\gamma^{\mu\beta}}{(k+\gamma)^{\mu\beta}} + \sum_{t=0}^{k-1}\left(\frac{\widehat{u}\beta^2}{(t+\gamma)^2} + \frac{\widehat{z}\beta^3}{(t+\gamma)^3}\right)\frac{(t+1+\gamma)^{\mu\beta}}{(k+\gamma)^{\mu\beta}} \\
&= \Psi_0 \frac{\gamma^{\mu\beta}}{(k+\gamma)^{\mu\beta}} + \frac{\widehat{u}\beta^2}{(k+\gamma)^{\mu\beta}}\sum_{t=0}^{k-1}\frac{(t+1+\gamma)^{\mu\beta}}{(t+\gamma)^2} + \frac{\widehat{z}\beta^3}{(k+\gamma)^{\mu\beta}}\sum_{t=0}^{k-1}\frac{(t+1+\gamma)^{\mu\beta}}{(t+\gamma)^3},
\end{aligned} \tag{4.69}$$

where the second inequality uses Lemma 4.5.21. By $1+x \leq e^x, \forall x \in \mathbb{R}$, we have: for $0 \leq t \leq k-1$,

$$\frac{(t+1+\gamma)^{\mu\beta}}{(t+\gamma)^{\mu\beta}} = \left(1+\frac{1}{t+\gamma}\right)^{\mu\beta} \leq \exp\left\{\frac{\mu\beta}{\gamma}\right\} \leq \sqrt{e}, \tag{4.70}$$

where the last inequality uses $\mu\beta/\gamma \leq \mu\overline{\alpha} \leq 0.5$. We use (4.70) in (4.69) to obtain: $\forall k \geq 1$,

$$\Psi_k \leq \Psi_0 \frac{\gamma^{\mu\beta}}{(k+\gamma)^{\mu\beta}} + \frac{\sqrt{e}\widehat{u}\beta^2}{(k+\gamma)^{\mu\beta}}\sum_{s=\gamma}^{k-1+\gamma} s^{\mu\beta-2} + \frac{\sqrt{e}\widehat{z}\beta^3}{(k+\gamma)^{\mu\beta}}\sum_{s=\gamma}^{k-1+\gamma} s^{\mu\beta-3}. \tag{4.71}$$

By $s^{\mu\beta-2} \leq \max\left\{\int_s^{s+1} x^{\mu\beta-2}dx, \int_{s-1}^{s} x^{\mu\beta-2}dx\right\}$, we have: if $\beta > 1/\mu$, then $\forall k \geq 1$,

$$\sum_{s=\gamma}^{k-1+\gamma} s^{\mu\beta-2} \leq \int_{\gamma-1}^{k+\gamma} x^{\mu\beta-2}dx \leq \frac{(k+\gamma)^{\mu\beta-1}}{\mu\beta-1}. \tag{4.72}$$

Likewise, by $s^{\mu\beta-3} \leq \max\left\{\int_s^{s+1} x^{\mu\beta-3}dx, \int_{s-1}^{s} x^{\mu\beta-3}dx\right\}$, we have: if $\beta > 2/\mu$, then $\forall k \geq 1$,

$$\sum_{s=\gamma}^{k-1+\gamma} s^{\mu\beta-3} \leq \int_{\gamma-1}^{k+\gamma} x^{\mu\beta-3}dx \leq \frac{(k+\gamma)^{\mu\beta-2}}{\mu\beta-2}. \tag{4.73}$$

Now, we apply (4.72) and (4.73) in (4.71) to obtain: $\forall k \geq 1$,

$$\Psi_k \leq \frac{\Psi_0\gamma^{\mu\beta}}{(k+\gamma)^{\mu\beta}} + \frac{\sqrt{e}\widehat{u}\beta^2}{(\mu\beta-1)(k+\gamma)} + \frac{\sqrt{e}\widehat{z}\beta^3}{(\mu\beta-2)(k+\gamma)^2}. \tag{4.74}$$

Using (4.74) and Lemma 4.5.20 in Lemma 4.5.13, we obtain: $\forall k \geq 1$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[F(\mathbf{x}_k^i) - F^*] \leq \frac{2(F(\overline{\mathbf{x}}_0)-F^*)}{(k/\gamma+1)^{\mu\beta}} + \frac{2\sqrt{e}L\widehat{u}\beta^2}{(\mu\beta-1)(k+\gamma)} + \frac{2\sqrt{e}L\widehat{z}\beta^3}{(\mu\beta-2)(k+\gamma)^2} + \frac{2\widehat{z}\beta^2}{(k+\gamma)^2}.$$

The proof follows by that $\frac{\widehat{z}\beta^2}{(k+\gamma)^2} \leq \frac{L\widehat{z}\beta^3}{(\mu\beta-2)(k+\gamma)^2}$ and by recalling the definitions of $\widehat{u}$ and $\widehat{z}$ given in (4.68). □

[6] For a sequence $\{s_k\}$, we adopt the convention $\prod_{k=x}^{y} s_k = 1$ if $y < x$.

Proof of Corollary 4.3.2. We derive the conditions under which the rate expression in Theorem 4.3.4 is network-independent. We first solve for the lower bound on k such that

$$\frac{L\nu_a^2\beta^2}{n(\mu\beta-1)(k+\gamma)} \geq \frac{L^2\widehat{x}\beta^3}{n(\mu\beta-2)(k+\gamma)^2},$$

which may be written equivalently as

$$k+\gamma \geq \frac{\mu\beta-1}{\mu\beta-2}\frac{L\widehat{x}\beta}{\nu_a^2}. \tag{4.75}$$

We suppose that $\|\nabla \mathbf{f}_0\|^2 = \mathcal{O}(n)$, $\beta = \theta/\mu$, where $\theta > 2$. Since $\overline{\alpha}L = \mathcal{O}\big(\frac{1-\lambda}{\lambda\kappa^{1/4}}\big)$, for $\overline{\alpha}$ defined in (4.31), we have

$$\widehat{x} = \mathcal{O}\left(\frac{\lambda^2 n\nu_a^2}{(1-\lambda)^3} + \frac{\lambda\kappa^{1/4}\nu_a^2}{1-\lambda} + \frac{\lambda^2 nL(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda)^2\kappa^{1/2}}\right),$$

where $\widehat{x}$ is defined in (4.63). Therefore, to make (4.75) hold, it suffices to let

$$k \gtrsim \frac{\lambda^2 n\kappa}{(1-\lambda)^3} + \frac{\lambda\kappa^{5/4}}{1-\lambda} + \frac{\lambda^2 n\kappa^{1/2}L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda)^2\nu_a^2}. \tag{4.76}$$

Next, we solve for the range of k such that for some $\delta \in [1,\theta)$, $(\frac{k}{\gamma}+1)^\theta \geq (\frac{k+1}{\kappa})^\delta$, i.e., $\frac{(k+\gamma)^\theta}{(k+1)^\delta} \geq \frac{\gamma^\theta}{\kappa^\delta}$. Since $\gamma > 1$, it suffices choose k such that

$$k \geq \gamma^{\frac{\theta}{\theta-\delta}}\kappa^{-\frac{\delta}{\theta-\delta}}. \tag{4.77}$$

We fix $\gamma = \max\{\frac{\theta}{\mu\overline{\alpha}}, \frac{8}{1-\lambda^2}\} \asymp \max\{\kappa, \frac{\lambda^2\kappa}{(1-\lambda)^2}, \frac{\lambda\kappa^{5/4}}{1-\lambda}, \frac{1}{1-\lambda}\}$. Using (4.76) and (4.77) in Theorem 4.3.4, we have

$$\frac{1}{n}\sum_{i=1}^n \big(F(\mathbf{x}_k^i)-F^*\big) = \mathcal{O}\left(\frac{\kappa^\delta\,(F(\overline{\mathbf{x}}_0)-F^*)}{k^\delta} + \frac{\kappa\nu_a^2}{n\mu k}\right),$$

if $k \gtrsim \max\{K_1, K_2\}$, where K_1 and K_2 are given by

$$K_1 = \frac{\lambda^2 n\kappa}{(1-\lambda)^3} + \frac{\lambda\kappa^{5/4}}{1-\lambda} + \frac{\lambda^2 n\kappa^{1/2}L(F(\overline{\mathbf{x}}_0)-F^*)}{(1-\lambda)^2\nu_a^2},$$

$$K_2 = \max\left\{\kappa, \frac{\lambda^2\kappa}{(1-\lambda)^2}, \frac{\lambda\kappa^{5/4}}{1-\lambda}, \frac{1}{1-\lambda}\right\}^{\frac{\theta}{\theta-\delta}}\kappa^{-\frac{\delta}{\theta-\delta}}.$$

The proof follows by setting $\delta = 2$ and $\theta = 6$ in the above. □

4.6 Conclusion

In this chapter, we study the convergence properties of the well-known `GT-DSGD` algorithm for decentralized smooth non-convex expected risk minimization problems. For both constant and decaying step-sizes, we comprehensively establish the conditions under which the performances of `GT-DSGD` are network topology-independent and match that of the centralized `SGD` algorithm for general non-convex problems and problems where the global PL condition is satisfied. In sharp contrast, the existing theory suggests that the performances of `GT-DSGD` are strictly worse than that of the centralized `SGD`.

Chapter 5

Decentralized Online Stochastic Non-Convex Optimization with Mean-Squared Smoothness

In this chapter, we study decentralized non-convex expected risk minimization problems with mean-squared smoothness. Inspired by the `GT-VR` framework proposed in Chapter 2, we propose `GT-HSGD`, a new single-loop decentralized variance-reduced stochastic gradient method, which achieves improved oracle complexity and practical implementation compared with the existing approaches. In particular, we show that `GT-HSGD` achieves an ϵ-accurate stationary point of the problem with a network topology-independent oracle complexity of $O(\epsilon^{-3})$ that matches the centralized optimal methods for this problem class, when the required error tolerance ϵ is small enough. We present numerical experiments to verify our main technical results.

5.1 Introduction

We consider n nodes, such as machines or edge devices, communicating over a decentralized network described by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \cdots, n\}$ is the set of node indices and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the collection of ordered pairs (i, j), $i, j \in \mathcal{V}$, such that node j sends information to node i. Each node i possesses a private local cost function $f_i : \mathbb{R}^p \to \mathbb{R}$ and the goal of the networked nodes is to solve, via local computation and communication, the following optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^p} F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}).$$

This canonical formulation is known as decentralized optimization [19, 21, 33, 34] that has emerged as a promising framework for large-scale data science and machine learning problems [2, 41]. Decentralized optimization is essential in scenarios where data is geographically distributed and/or centralized data processing is infeasible due to communication and computation overhead or data privacy concerns. In this chapter, we

focus on an *online and non-convex* setting. In particular, we assume that each local cost f_i is *non-convex* and each node i only accesses f_i by querying a local *stochastic first-order oracle (SFO)* [128] that returns a stochastic gradient, i.e., a noisy version of the exact gradient, at the queried point. As a concrete example of practical interest, the SFO mechanism applies to many online learning and expected risk minimization problems where the noise in SFO lies in the uncertainty of sampling from the underlying streaming data received at each node [19, 21]. We are interested in the oracle complexity, i.e., the total number of queries to SFO required at each node, to find an ϵ-accurate first-order stationary point $\mathbf{x}^*$ of the global cost F such that $\mathbb{E}[\|\nabla F(\mathbf{x}^*)\|] \leq \epsilon$.

5.1.1 Related work

We now briefly review the literature of decentralized non-convex optimization with SFO, which has been widely studied recently. Perhaps the most well-known approach is the decentralized stochastic gradient descent (DSGD) and its variants [2, 19, 21, 43, 155], which combine average consensus and a local stochastic gradient step. Although being simple and effective, DSGD is known to have difficulties in handling heterogeneous data [31]. Recent works [3,4,148,150] achieve robustness to heterogeneous environments by leveraging certain decentralized bias-correction techniques such as EXTRA (type) [20, 68, 156], gradient tracking [14, 54–56,65,67,157], and primal-dual principles [66,69,73,74]. Built on top of these bias-correction techniques, very recent works [135] and [158] propose D-GET and D-SPIDER-SFO respectively that further incorporate online SARAH/SPIDER-type variance reduction schemes [48–50] to achieve lower oracle complexities, when the SFO satisfies a mean-squared smoothness property. Finally, we note that the family of decentralized variance reduced methods has been significantly enriched recently, see, for instance, [22, 23, 31, 120, 141, 142, 159–161]; however, these approaches are explicitly designed for empirical minimization where each local cost f_i is decomposed as a finite-sum of component functions, i.e., $f_i = \frac{1}{m}\sum_{r=1}^{m} f_{i,r}$; it is therefore unclear whether these algorithms can be adapted to the online SFO setting, which is the focus of this chapter.

5.1.2 Main contributions

In this chapter, we propose GT-HSGD, a novel online variance reduced method for decentralized non-convex optimization with stochastic first-order oracles (SFO). To achieve fast and robust performance, the GT-HSGD algorithm is built upon global gradient tracking [55, 65] and a local hybrid stochastic gradient estimator [59, 162, 163] that can be considered as a convex combination of the vanilla stochastic gradient returned by the SFO and a SARAH-type variance-reduced stochastic gradient [58]. In the following, we emphasize the key advantages of GT-HSGD compared with the existing decentralized online (variance-reduced) approaches, from

both theoretical and practical aspects.

- **Improved oracle complexity**. A comparison of the oracle complexity of `GT-HSGD` with related algorithms is provided in Table 5.1, from which we have the following important observations. First of all, the oracle complexity of `GT-HSGD` is lower than that of `DSGD`, `D2`, `GT-DSGD` and `D-PD-SGD`, which are decentralized online algorithms without variance reduction; however, `GT-HSGD` imposes on the `SFO` an additional mean-squared smoothness (MSS) assumption that is required by all online variance-reduced techniques in the literature [48–50, 59, 134, 135, 149, 158, 162, 163]. Secondly, `GT-HSGD` further achieves a lower oracle complexity than the existing decentralized online variance-reduced methods `D-GET` [135] and `D-SPIDER-SFO` [158], especially in a regime where the required error tolerance ϵ and the network spectral gap $(1-\lambda)$ are relatively small.[1]. Moreover, when ϵ is small enough such that $\epsilon \lesssim \min\left\{\lambda^{-4}(1-\lambda)^3 n^{-1}, \lambda^{-1}(1-\lambda)^{1.5} n^{-1}\right\}$, it can be verified that the oracle complexity of `GT-HSGD` reduces to $O(n^{-1}\epsilon^{-3})$, independent of the network topology, and `GT-HSGD` achieves a linear speedup, in terms of the scaling with the network size n, compared with the centralized optimal online variance-reduced approaches that operate on a single node [48–50, 59, 134, 162]; see Section 5.3 for a detailed discussion. In sharp contrast, the speedup of `D-GET` [135] and `D-SPIDER-SFO` [158] is not clear compared with the aforementioned centralized optimal methods even if the network is fully connected, i.e., $\lambda = 0$.

- **More practical implementation**. Both `D-GET` [135] and `D-SPIDER-SFO` [158] are double-loop algorithms that require very large minibatch sizes. In particular, during each inner loop they execute a fixed number of minibatch stochastic gradient type iterations with $O(\epsilon^{-1})$ oracle queries per update per node, while at every outer loop they obtain a stochastic gradient with mega minibatch size by $O(\epsilon^{-2})$ oracle queries at each node. Clearly, querying the oracles exceedingly, i.e., obtaining a large amount of samples, at each node and every iteration in online steaming data scenarios substantially jeopardizes the actual wall-clock time. This is because the next iteration cannot be performed until all nodes complete the sampling process. Moreover, the double-loop implementation may incur periodic network synchronizations. These issues are especially significant when the working environments of the nodes are heterogeneous. Conversely, the proposed `GT-HSGD` is a single-loop algorithm with $O(1)$ oracle queries per update and only requires a large minibatch size with $O(\epsilon^{-1})$ oracle queries once in the *initialization phase*, i.e., before the update recursion is executed; see Algorithm 6 and Corollary 5.3.1.

[1] A small network spectral gap $(1-\lambda)$ implies that the connectivity of the network is weak.

Table 5.1: A comparison of the oracle complexity of decentralized online stochastic gradient methods. The oracle complexity is in terms of the total number of queries to SFO required *at each node* to obtain an ϵ-accurate stationary point $\mathbf{x}^*$ of the global cost F such that $\mathbb{E}[\|\nabla F(\mathbf{x}^*)\|] \leq \epsilon$. In the table, n is the number of the nodes and $(1-\lambda) \in (0,1]$ is the spectral gap of the weight matrix associated with the network. We note that the complexity of D2 and D-SPIDER-SFO also depends on the smallest eigenvalue λ_n of the weight matrix; however, since λ_n is less sensitive to the network topology, we omit the dependence of λ_n in the table for conciseness. The MSS column indicates whether the algorithm in question requires the mean-squared smoothness assumption on the SFO. Finally, we emphasize that DSGD requires bounded heterogeneity such that $\sup_{\mathbf{x}} \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq \zeta^2$, for some $\zeta \in \mathbb{R}^+$, while other algorithms in the table do not need this assumption.

Algorithm	**Oracle Complexity**	**MSS**	**Remarks**
DSGD [2]	$O\left(\max\left\{\frac{1}{n\epsilon^4}, \frac{\lambda^2 n}{(1-\lambda)^2\epsilon^2}\right\}\right)$	✗	bounded heterogeneity
D2 [3]	$O\left(\max\left\{\frac{1}{n\epsilon^4}, \frac{n}{(1-\lambda)^b\epsilon^2}\right\}\right)$	✗	$b \in \mathbb{R}^+$ is not explicitly shown in [3]
GT-DSGD [4]	$O\left(\max\left\{\frac{1}{n\epsilon^4}, \frac{\lambda^2 n}{(1-\lambda)^3\epsilon^2}\right\}\right)$	✗	
D-PD-SGD [148]	$O\left(\max\left\{\frac{1}{n\epsilon^4}, \frac{n}{(1-\lambda)^c\epsilon^2}\right\}\right)$	✗	$c \in \mathbb{R}^+$ is not explicitly shown in [148]
D-GET [135]	$O\left(\frac{1}{(1-\lambda)^d\epsilon^3}\right)$	✓	$d \in \mathbb{R}^+$ is not explicitly shown in [135]
D-SPIDER-SFO [158]	$O\left(\frac{1}{(1-\lambda)^h\epsilon^3}\right)$	✓	$h \in \mathbb{R}^+$ is not explicitly shown in [158]
GT-HSGD **(this work)**	$O\left(\max\left\{\frac{1}{n\epsilon^3}, \frac{\lambda^4}{(1-\lambda)^3\epsilon^2}, \frac{\lambda^{1.5}n^{0.5}}{(1-\lambda)^{2.25}\epsilon^{1.5}}\right\}\right)$	✓	

The rest of this chapter is organized as follows. In Section 5.2, we state the problem formulation and develop the proposed GT-HSGD algorithm. Section 5.3 presents the main convergence results of GT-HSGD and their implications. Section 5.4 provides numerical experiments to illustrate our theoretical claims. Section 5.5 outlines the convergence analysis of GT-HSGD, while the detailed proofs and derivations are provided in Section 5.6. Section 5.7 concludes the chapter.

We adopt the following notations throughout the chapter. We use lowercase bold letters to denote vectors and uppercase bold letters to denote matrices. The ceiling function is denoted as $\lceil\cdot\rceil$. The matrix $\mathbf{I}_d$ represents the $d\times d$ identity; $\mathbf{1}_d$ and $\mathbf{0}_d$ are the d-dimensional column vectors of all ones and zeros, respectively. We denote $[\mathbf{x}]_i$ as the i-th entry of a vector $\mathbf{x}$. The Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A}\otimes\mathbf{B}$. We use $\|\cdot\|$ to denote the Euclidean norm of a vector or the spectral norm of a matrix. We use $\sigma(\cdot)$ to denote the σ-algebra generated by the sets and/or random vectors in its argument.

5.2 Problem setup and the GT-HSGD algorithm

In this section, we introduce the mathematical model of the stochastic first-order oracle (SFO) at each node and the communication network. Based on these formulations, we develop the proposed GT-HSGD algorithm.

5.2.1 Optimization and network model

We work with a rich enough probability space $\{\Omega, \mathbb{P}, \mathcal{F}\}$. We consider decentralized recursive algorithms of interest that generate a sequence of estimates $\{\mathbf{x}_t^i\}_{t\geq 0}$ of the first-order stationary points of F at each node i, where $\mathbf{x}_0^i$ is assumed constant. At each iteration t, each node i observes a random vector $\boldsymbol{\xi}_t^i$ in $\mathbb{R}^q$, which, for instance, may be considered as noise or as an online data sample. We then introduce the natural filtration (an increasing family of sub-σ-algebras of $\mathcal{F}$) induced by these random vectors observed sequentially by the networked nodes:

$$\begin{aligned} &\mathcal{F}_0 := \{\Omega, \phi\}, \\ &\mathcal{F}_t := \sigma\left(\left\{\boldsymbol{\xi}_0^i, \boldsymbol{\xi}_1^i, \cdots, \boldsymbol{\xi}_{t-1}^i : i \in \mathcal{V}\right\}\right), \qquad \forall t \geq 1, \end{aligned} \tag{5.1}$$

where ϕ is the empty set. We are now ready to define the `SFO` mechanism in the following. At each iteration t, each node i, given an input random vector $\mathbf{x} \in \mathbb{R}^p$ that is $\mathcal{F}_t$-measurable, is able to query the local `SFO` to obtain a stochastic gradient of the form $\mathbf{g}_i(\mathbf{x}, \boldsymbol{\xi}_t^i)$, where $\mathbf{g}_i : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^p$ is a Borel measurable function. We assume that the `SFO` satisfies the following four properties.

Assumption 5.2.1 (**Oracle**)**.** For any $\mathcal{F}_t$-measurable random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, we have the following: $\forall i \in \mathcal{V}$, $\forall t \geq 0$,

- $\mathbb{E}\left[\mathbf{g}_i(\mathbf{x}, \boldsymbol{\xi}_t^i)|\mathcal{F}_t\right] = \nabla f_i(\mathbf{x})$;
- $\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{x}, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x})\|^2\right] \leq \nu_i^2$, $\overline{\nu}^2 := \frac{1}{n}\sum_{i=1}^n \nu_i^2$;
- the family $\left\{\boldsymbol{\xi}_t^i : \forall t \geq 0, i \in \mathcal{V}\right\}$ of random vectors is independent;
- $\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{x}, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{y}, \boldsymbol{\xi}_t^i)\|^2\right] \leq L^2\mathbb{E}\left[\|\mathbf{x} - \mathbf{y}\|^2\right]$.

The first three properties above are standard and commonly used to establish the convergence of decentralized stochastic gradient methods. They however do not explicitly impose any structures on the stochastic gradient mapping $\mathbf{g}_i$ other than the measurability. On the other hand, the last property, the mean-squared smoothness, roughly speaking, requires that $\mathbf{g}_i$ is L-smooth on average with respect to the input arguments $\mathbf{x}$ and $\mathbf{y}$. As a simple example, Assumption 5.2.1 holds if $f_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{Q}_i\mathbf{x}$ and $\mathbf{g}_i(\mathbf{x}, \boldsymbol{\xi}_i) = \mathbf{Q}_i\mathbf{x} + \boldsymbol{\xi}_i$, where $\mathbf{Q}_i$ is a constant matrix and $\boldsymbol{\xi}_i$ has zero mean and finite second moment. We further note that the mean-squared smoothness of each $\mathbf{g}_i$ implies, by Jensen's inequality, that each f_i is L-smooth, i.e., $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, and consequently the global function F is also L-smooth.

In addition, we make the following assumptions on F and the communication network $\mathcal{G}$.

Assumption 5.2.2 (**Global Function**)**.** F is bounded below, i.e., $F^* := \inf_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) > -\infty$.

Assumption 5.2.3 (Communication Network). The directed network $\mathcal{G}$ admits a primitive and doubly-stochastic weight matrix $\underline{\mathbf{W}} = \{\underline{w}_{ij}\} \in \mathbb{R}^{n\times n}$. Hence, $\underline{\mathbf{W}}\mathbf{1}_n = \underline{\mathbf{W}}^\top \mathbf{1}_n = \mathbf{1}_n$ and $\lambda := \|\underline{\mathbf{W}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\| \in [0,1)$.

The weight matrix $\underline{\mathbf{W}}$ that satisfies Assumption 5.2.3 may be designed for strongly-connected weight-balanced directed graphs (and thus for arbitrary connected undirected graphs). For example, the family of directed exponential graphs is weight-balanced and plays a key role in decentralized training [41]. We note that λ is known as the second largest singular value of $\underline{\mathbf{W}}$ and measures the algebraic connectivity of the graph, i.e., a smaller value of λ roughly means a better connectivity. We note that several existing approaches require strictly stronger assumptions on $\underline{\mathbf{W}}$. For instance, `D2` [3] and `D-PD-SGD` [148] require $\underline{\mathbf{W}}$ to be symmetric and hence are restricted to undirected networks.

5.2.2 Algorithm development

We now describe the proposed `GT-HSGD` algorithm and provide an intuitive construction. Recall that $\mathbf{x}_t^i$ is the estimate of an stationary point of the global cost F at node i and iteration t. Let $\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i)$ and $\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)$ be the corresponding stochastic gradients returned by the local `SFO` queried at $\mathbf{x}_t^i$ and $\mathbf{x}_{t-1}^i$ respectively. Motivated by the strong performance of recently introduced decentralized methods that combine gradient tracking and various variance reduction schemes for finite-sum problems [120, 135, 141, 160], we seek similar variance reduction for decentralized online problems with `SFO`. In particular, we focus on the following *local* hybrid variance reduced stochastic gradient estimator $\mathbf{v}_t^i$ introduced in [59, 162, 163] for centralized online problems: $\forall t \geq 1$,

$$\mathbf{v}_t^i = \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) + (1-\beta)\big(\mathbf{v}_{t-1}^i - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)\big), \tag{5.2}$$

for some applicable weight parameter $\beta \in [0,1]$. This local gradient estimator $\mathbf{v}_t^i$ is fused, via a gradient tracking mechanism [55, 65], over the network to update the global gradient tracker $\mathbf{y}_t^i$, which is subsequently used as the descent direction in the $\mathbf{x}_t^i$-update. The complete description of `GT-HSGD` is provided in Algorithm 6. We note that the update (5.2) of $\mathbf{v}_t^i$ may be equivalently written as

$$\mathbf{v}_t^i = \beta \cdot \underbrace{\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i)}_{\texttt{Stochastic gradient}} + (1-\beta)\cdot \underbrace{\big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \mathbf{v}_{t-1}^i\big)}_{\texttt{SARAH}},$$

which is a convex combination of the local vanilla stochastic gradient returned by the `SFO` and a `SARAH`-type [49, 50, 58] gradient estimator. This discussion leads to the fact that `GT-HSGD` reduces to `GT-DSGD` [4, 67, 150] when $\beta = 1$, and becomes the inner loop of `GT-SARAH` [141] when $\beta = 0$. However, our convergence analysis shows that `GT-HSGD` achieves its best oracle complexity and outperforms the existing decentralized online variance-reduced approaches [135, 158] with a weight parameter $\beta \in (0,1)$. It is then clear that neither

Algorithm 6 `GT-HSGD` at each node i

Require: $\mathbf{x}_0^i = \overline{\mathbf{x}}_0$; α; β; b_0; $\mathbf{y}_0^i = \mathbf{0}_p$; $\mathbf{v}_{-1}^i = \mathbf{0}_p$; T.

Sample $\{\boldsymbol{\xi}_{0,r}^i\}_{r=1}^{b_0}$ and $\mathbf{v}_0^i = \frac{1}{b_0}\sum_{r=1}^{b_0} \mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i)$;

$\mathbf{y}_1^i = \sum_{j=1}^n \underline{w}_{ij}\big(\mathbf{y}_0^j + \mathbf{v}_0^j - \mathbf{v}_{-1}^j\big)$;

$\mathbf{x}_1^i = \sum_{j=1}^n \underline{w}_{ij}\big(\mathbf{x}_0^j - \alpha \mathbf{y}_1^j\big)$;

for $t = 1, 2, \cdots, T-1$ **do**

  Sample $\boldsymbol{\xi}_t^i$;

  $\mathbf{v}_t^i = \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) + (1-\beta)\big(\mathbf{v}_{t-1}^i - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)\big)$.

  $\mathbf{y}_{t+1}^i = \sum_{j=1}^n \underline{w}_{ij}\big(\mathbf{y}_t^j + \mathbf{v}_t^j - \mathbf{v}_{t-1}^j\big)$;

  $\mathbf{x}_{t+1}^i = \sum_{j=1}^n \underline{w}_{ij}\big(\mathbf{x}_t^j - \alpha \mathbf{y}_{t+1}^j\big)$;

end for

return $\widetilde{\mathbf{x}}_T$ selected uniformly at random from $\{\mathbf{x}_t^i\}_{0\le t\le T}^{i\in\mathcal{V}}$.

`GT-DSGD` nor the inner loop of `GT-SARAH`, on their own, are able to outperform the proposed approach, making `GT-HSGD` a non-trivial algorithmic design for this problem class.

Remark 5.2.1. Clearly, each $\mathbf{v}_t^i$ is a conditionally biased estimator of $\nabla f_i(\mathbf{x}_t^i)$, i.e., $\mathbb{E}[\mathbf{v}_t^i|\mathcal{F}_t] \neq \nabla f_i(\mathbf{x}_t^i)$ in general. However, it can be shown that $\mathbb{E}[\mathbf{v}_t^i] = \mathbb{E}[\nabla f_i(\mathbf{x}_t^i)]$, meaning that $\mathbf{v}_t^i$ serves as a surrogate for the underlying exact gradient in the sense of total expectation.

5.3 Main results

In this section, we present the main convergence results of `GT-HSGD` in this chapter and discuss their salient features. The formal convergence analysis is deferred to Section 5.5.

Theorem 5.3.1. *If the weight parameter* $\beta = \frac{48L^2\alpha^2}{n}$ *and the step-size* α *is chosen as*

$$0 < \alpha < \min\left\{\frac{(1-\lambda^2)^2}{90\lambda^2}, \frac{\sqrt{n(1-\lambda)}}{26\lambda}, \frac{1}{4\sqrt{3}}\right\}\frac{1}{L},$$

then the output $\widetilde{\mathbf{x}}_T$ *of* `GT-HSGD` *satisfies:* $\forall T \geq 2$,

$$\mathbb{E}\big[\|\nabla F(\widetilde{\mathbf{x}}_T)\|^2\big] \leq \frac{4(F(\overline{\mathbf{x}}_0) - F^*)}{\alpha T} + \frac{8\beta\overline{\nu}^2}{n} + \frac{4\overline{\nu}^2}{\beta b_0 n T} + \frac{64\lambda^4\|\nabla \mathbf{f}(\mathbf{x}_0)\|^2}{(1-\lambda^2)^3 n T} + \frac{96\lambda^2\overline{\nu}^2}{(1-\lambda^2)^3 b_0 T} + \frac{256\lambda^2\beta^2\overline{\nu}^2}{(1-\lambda^2)^3},$$

where $\|\nabla \mathbf{f}(\mathbf{x}_0)\|^2 = \sum_{i=1}^n \|\nabla f_i(\overline{\mathbf{x}}_0)\|^2$

Remark 5.3.1. Theorem 5.3.1 holds for `GT-HSGD` with arbitrary initial minibatch size $b_0 \geq 1$.

Theorem 5.3.1 establishes a non-asymptotic bound, with no hidden constants, on the mean-squared stationary gap of `GT-HSGD` over any finite time horizon T.

Remark 5.3.2 (**Transient and steady-state performance over infinite time horizon**). If α and β are chosen according to Theorem 5.3.1, the mean-squared stationary gap $\mathbb{E}\left[\|\nabla F(\widetilde{\mathbf{x}}_T)\|^2\right]$ of `GT-HSGD` decays sublinearly at a rate of $O(1/T)$ up to a steady-state error (SSE) such that

$$\limsup_{T\to\infty} \mathbb{E}\left[\|\nabla F(\widetilde{\mathbf{x}}_T)\|^2\right] \leq \frac{8\beta\overline{\nu}^2}{n} + \frac{256\lambda^2\beta^2\overline{\nu}^2}{(1-\lambda^2)^3}. \tag{5.3}$$

In view of (5.3), the SSE of `GT-HSGD` is bounded by the sum of two terms: (i) the first term is in the order of $O(\beta)$ and the division by n demonstrates the benefit of increasing the network size[2]; (ii) the second term is in the order of $O(\beta^2)$ and reveals the impact of the spectral gap $(1-\lambda)$ of the network topology. Clearly, the SSE can be made arbitrarily small by choosing small enough β and α. Moreover, since the spectral gap $(1-\lambda)$ only appears in a higher order term of β in (5.3), its impact reduces as β becomes smaller, i.e., as we require a smaller SSE.

The following corollary is concerned with the finite-time convergence rate of `GT-HSGD` with specific choices of the algorithmic parameters α, β, and b_0.

Corollary 5.3.1. *Setting* $\alpha = \frac{n^{2/3}}{8LT^{1/3}}$, $\beta = \frac{3n^{1/3}}{4T^{2/3}}$, *and* $b_0 = \lceil \frac{T^{1/3}}{n^{2/3}} \rceil$ *in Theorem 5.3.1, we have*

$$\mathbb{E}\left[\|\nabla F(\widetilde{\mathbf{x}}_T)\|^2\right] \leq \frac{32L(F(\overline{\mathbf{x}}_0)-F^*)+12\overline{\nu}^2}{(nT)^{2/3}} + \frac{64\lambda^4\|\nabla \mathbf{f}(\mathbf{x}_0)\|^2}{(1-\lambda^2)^3 nT} + \frac{240\lambda^2 n^{2/3}\overline{\nu}^2}{(1-\lambda^2)^3 T^{4/3}},$$

for all

$$T \geq \max\left\{\frac{1424\lambda^6 n^2}{(1-\lambda^2)^6}, \frac{35\lambda^3 n^{0.5}}{(1-\lambda)^{1.5}}\right\}.$$

As a consequence, `GT-HSGD` *achieves an* ϵ*-accurate stationary point* $\mathbf{x}^*$ *of the global cost* F *such that* $\mathbb{E}[\|\nabla F(\mathbf{x}^*)\|] \leq \epsilon$ *with*

$$\mathcal{H} = O\left(\max\left\{\mathcal{H}_{opt}, \mathcal{H}_{net}\right\}\right)$$

iterations[3]*, where* $\mathcal{H}_{opt}$ *and* $\mathcal{H}_{net}$ *are given respectively by*

$$\mathcal{H}_{opt} = \frac{(L(F(\overline{\mathbf{x}}_0)-F^*)+\overline{\nu}^2)^{1.5}}{n\epsilon^3},$$
$$\mathcal{H}_{net} = \max\left\{\frac{\lambda^4\|\nabla \mathbf{f}(\mathbf{x}_0)\|^2}{(1-\lambda^2)^3 n\epsilon^2}, \frac{\lambda^{1.5} n^{0.5}\overline{\nu}^{1.5}}{(1-\lambda^2)^{2.25}\epsilon^{1.5}}\right\}.$$

The resulting total number of oracle queries at each node is thus $\lceil \mathcal{H} + \mathcal{H}^{1/3} n^{-2/3} \rceil$.

Remark 5.3.3. Since $\mathcal{H}^{1/3} n^{-2/3}$ is much smaller than $\mathcal{H}$, we treat the oracle complexity of `GT-HSGD` as $\mathcal{H}$ for the ease of exposition in Table 5.1 and the following discussion.

[2]Since `GT-HSGD` computes $O(n)$ stochastic gradients in parallel per iteration across the nodes, the network size n can be interpreted as the minibatch size of `GT-HSGD`.

[3]The $O(\cdot)$ notation here does not absorb any problem parameters, i.e., it only hides universal constants.

An important implication of Corollary 5.3.1 is given in the following.

Remark 5.3.4 (**A regime for network topology-independent oracle complexity and linear speedup**). According to Corollary 5.3.1, the oracle complexity of `GT-HSGD` at each node is bounded by the maximum of two terms: (i) the first term $\mathcal{H}_{opt}$ is independent of the network topology and, more importantly, is n times smaller than the oracle complexity of the optimal centralized online variance-reduced methods that execute on a single node for this problem class [48–50, 59, 162]; (ii) the second term $\mathcal{H}_{net}$ depends on the network spectral gap $1-\lambda$ and is in the lower order of $1/\epsilon$. These two observations lead to the interesting fact that the oracle complexity of `GT-HSGD` becomes independent of the network topology, i.e., $\mathcal{H}_{opt}$ dominates $\mathcal{H}_{net}$, if the required error tolerance ϵ is small enough such that[4] $\epsilon \lesssim \min\left\{\lambda^{-4}(1-\lambda)^3 n^{-1}, \lambda^{-1}(1-\lambda)^{1.5} n^{-1}\right\}$. In this regime, `GT-HSGD` thus achieves a network topology-independent oracle complexity $\mathcal{H}_{opt} = O(n^{-1}\epsilon^{-3})$, exhibiting a linear speed up compared with the aforementioned centralized optimal algorithms [48–50, 59, 134, 162], in the sense that the total number of oracle queries required to achieve an ϵ-accurate stationary point at each node is reduced by a factor of $1/n$.

Remark 5.3.5. The small error tolerance regime in the above discussion corresponds to a large number of oracle queries, which translates to the scenario where the required total number of iterations T is large. Note that a large T further implies that the step-size α and the weight parameter β are small; see the expression of α and β in Corollary 5.3.1.

5.4 Numerical Experiments

In this section, we illustrate our theoretical results on the convergence of the proposed `GT-HSGD` algorithm with the help of numerical experiments. The basic setup is given in the following.

- **Model.** We consider a non-convex logistic regression model [137], where the decentralized non-convex optimization problem of interest takes the form $\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}) + r(\mathbf{x})$, such that

$$f_i(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^m \log\left[1 + e^{-\langle \mathbf{x}, \boldsymbol{\theta}_{ij}\rangle l_{ij}}\right]$$

and

$$r(\mathbf{x}) = R\sum_{k=1}^p \frac{[\mathbf{x}]_k^2}{1+[\mathbf{x}]_k^2},$$

where $\boldsymbol{\theta}_{i,j}$ is the feature vector, $l_{i,j} \in \{-1, +1\}$ is the corresponding binary label, and $r(\mathbf{x})$ is a non-convex regularizer. To simulate the online `SFO` setting described in Section 5.2, each node i is only able to *sample with replacement* from its local data $\{\boldsymbol{\theta}_{i,j}, l_{i,j}\}_{j=1}^m$ and compute the corresponding

[4] This boundary condition follows from basic algebraic manipulations.

Table 5.2: Datasets used in numerical experiments, all available at https://www.openml.org/.

Dataset	**train** (nm)	**dimension** (p)
a9a	48,840	123
covertype	100,000	54
KDD98	75,000	477
MiniBooNE	100,000	11

(minibatch) stochastic gradient. Throughout all experiments, we set the number of the nodes to $n = 20$ and the regularization parameter to $R = 10^{-4}$.

- **Data.** To test the performance of the applicable algorithms, we distribute the a9a, covertype, KDD98, MiniBooNE datasets uniformly over the nodes and normalize the feature vectors such that $\|\boldsymbol{\theta}_{i,j}\| = 1, \forall i, j$. The statistics of these datasets are provided in Table 5.2.

- **Network topology.** We consider the following network topologies: the undirected ring graph, the undirected and directed exponential graphs, and the complete graph; see [2, 14, 27, 41] for detailed configurations of these graphs. For all graphs, the associated doubly stochastic weights are set to be equal. The resulting second largest singular value λ of the weight matrices are $0.98, 0.75, 0.67, 0$, respectively, demonstrating a significant difference in the algebraic connectivity of these graphs.

- **Performance measure.** We measure the performance of the decentralized algorithms in question by the decrease of the global cost function value $F(\overline{\mathbf{x}})$, to which we refer as loss, versus epochs, where $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ with $\mathbf{x}_i$ being the model at node i and each epoch contains m stochastic gradient computations at each node.

5.4.1 Comparison with the existing decentralized stochastic gradient methods

We conduct a performance comparison of `GT-HSGD` with `GT-DSGD` [4, 67, 150], `D-GET` [135], and `D-SPIDER-SFO` [158] over the undirected exponential graph of 20 nodes. Note that we use `GT-DSGD` to represent methods that do not incorporate online variance reduction techniques, since it in general matches or outperforms `DSGD` [2] and has a similar performance with `D2` [3] and `D-PD-SGD` [148].

We set the parameters of `GT-HSGD`, `GT-DSGD`, `D-GET`, and `D-SPIDER-SFO` according to the following procedures. *First*, we find a very large step-size candidate set for each algorithm in comparison. *Second*, we choose the minibatch size candidate set for all algorithms as $\mathcal{B} := \{1, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$: the minibatch size of `GT-DSGD`, the minibatch size of `GT-HSGD` at $t = 0$, the minibatch size of `D-GET` and `D-SPIDER-SFO` at inner- and outer-loop are all chosen from $\mathcal{B}$. *Third*, for `D-GET` and `D-SPIDER-SFO`, we choose the inner-loop length candidate set as $\{\frac{m}{20b}, \frac{m}{19b}, \cdots, \frac{m}{b}, \frac{2m}{b}, \cdots, \frac{20m}{b}\}$, where m is the local data size

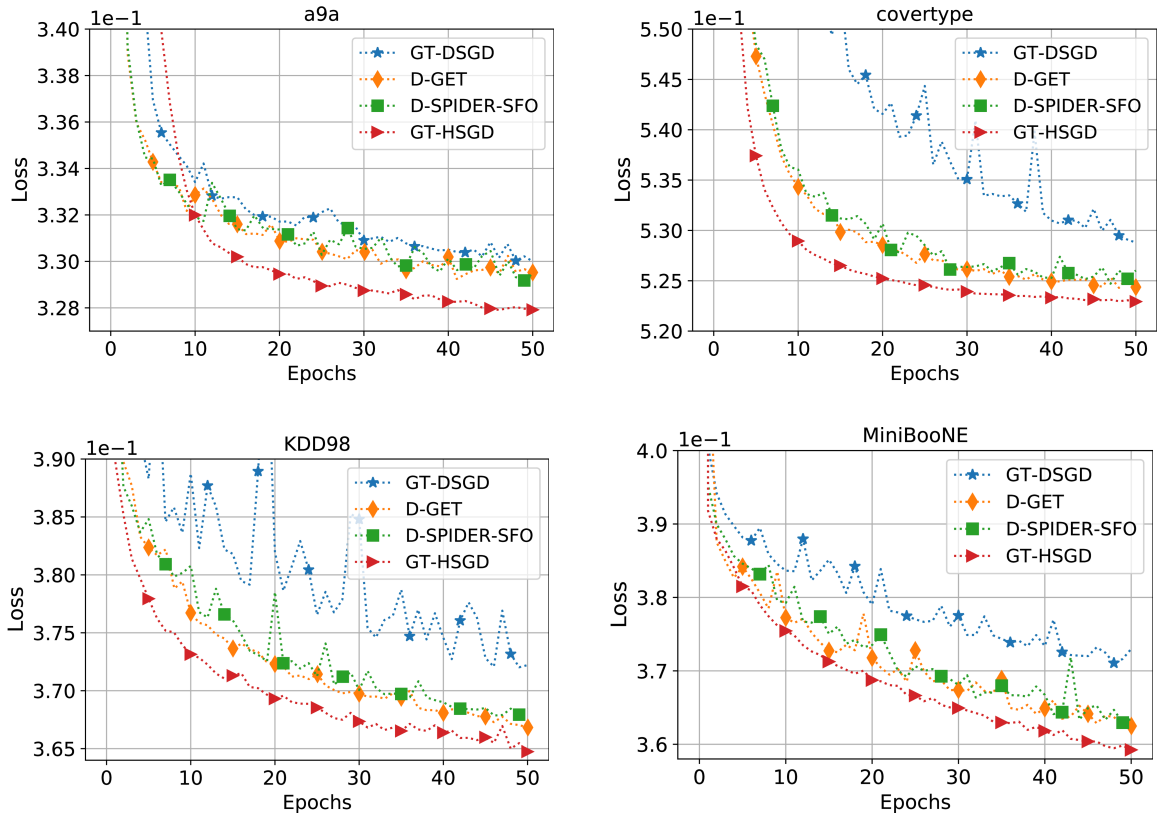

Figure 5.1: A comparison of GT-HSGD with other decentralized online stochastic gradient algorithms over the undirected exponential graph of 20 nodes on the a9a, covertype, KDD98, and MiniBooNE datasets.

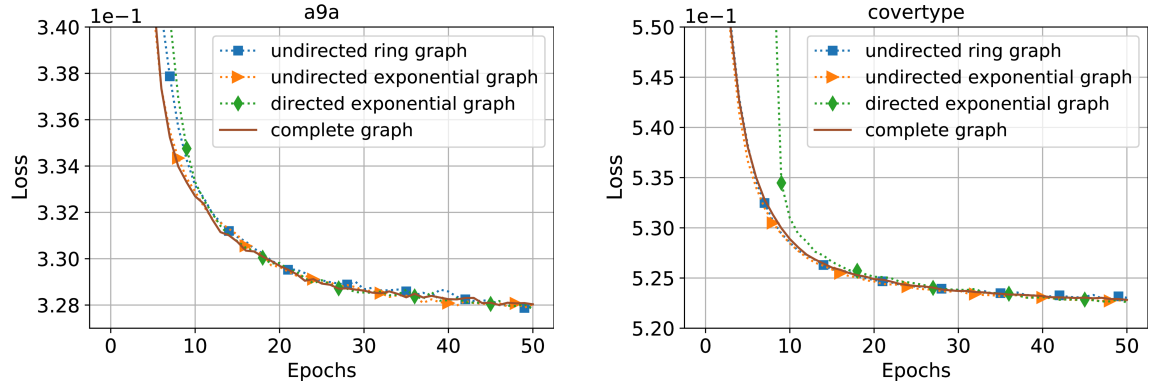

Figure 5.2: Convergence of GT-HSGD over different network topologies on the a9a and covertype datasets.

and b is the minibatch size at the inner-loop. *Fourth*, we iterate over all combinations of parameters for each algorithm to find its best performance. In particular, we find that the best performance of GT-HSGD is attained with a small β and a relatively large α as Corollary 5.3.1 suggests.

The experimental results are provided in Fig. 5.1, where we observe that GT-HSGD achieves faster rate than other algorithms in comparison on those four datasets. This observation is coherent with our main results that GT-HSGD achieves a lower oracle complexity than the existing approaches; see Table 5.1.

5.4.2 Topology-independent rate of GT-HSGD

We test the performance of GT-HSGD over different network topologies. In particular, we follow the procedures described in Section 5.4.1 to find the best set of parameters for GT-HSGD over the *complete graph* and then use this parameter set for other graphs. The corresponding experimental results are presented in Fig. 5.2. Clearly, it can be observed that when the number of iterations is large enough, that is to say, the required error tolerance is small enough, the convergence rate of GT-HSGD is not affected by the underlying network topology. This interesting phenomenon is consistent with our convergence theory; see Corollary 5.3.1 and the related discussion in Section 5.3.

5.5 Outline of the convergence analysis

In this section, we outlines the proof of Theorem 5.3.1, while the detailed proofs are provided in the Appendix. We let Assumptions 5.2.1-5.2.3 hold without explicitly stating them. For the ease of exposition, we write the $\mathbf{x}_t$- and $\mathbf{y}_t$-update of GT-HSGD in the following equivalent matrix form: $\forall t \geq 0$,

$$\mathbf{y}_{t+1} = \mathbf{W}\left(\mathbf{y}_t + \mathbf{v}_t - \mathbf{v}_{t-1}\right), \tag{5.4a}$$

$$\mathbf{x}_{t+1} = \mathbf{W}\left(\mathbf{x}_t - \alpha \mathbf{y}_{t+1}\right), \tag{5.4b}$$

where $\mathbf{W} := \underline{\mathbf{W}} \otimes \mathbf{I}_p$ and $\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t$ are square-integrable random vectors in $\mathbb{R}^{np}$ that respectively concatenate the local estimates $\{\mathbf{x}_t^i\}_{i=1}^n$ of a stationary point of F, gradient trackers $\{\mathbf{y}_t^i\}_{i=1}^n$, stochastic gradient estimators $\{\mathbf{v}_t^i\}_{i=1}^n$. It is straightforward to verify that $\mathbf{x}_t$ and $\mathbf{y}_t$ are $\mathcal{F}_t$-measurable while $\mathbf{v}_t$ is $\mathcal{F}_{t+1}$-measurable for all $t \geq 0$. For convenience, we also denote

$$\nabla \mathbf{f}(\mathbf{x}_t) := \left[\nabla f_1(\mathbf{x}_t^1)^\top, \cdots, \nabla f_n(\mathbf{x}_t^n)^\top\right]^\top$$

and introduce the following quantities:

$$\begin{aligned}
\mathbf{J} &:= \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right) \otimes \mathbf{I}_p \\
\overline{\mathbf{x}}_t &:= \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{x}_t, \\
\overline{\mathbf{y}}_t &:= \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{y}_t, \\
\overline{\mathbf{v}}_t &:= \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\mathbf{v}_t, \\
\overline{\nabla \mathbf{f}}(\mathbf{x}_t) &:= \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\nabla \mathbf{f}(\mathbf{x}_t).
\end{aligned}$$

In the following lemma, we enlist several well-known results in the context of gradient tracking-based algorithms for decentralized stochastic optimization, whose proofs may be found in [24, 55, 56, 67].

Lemma 5.5.1. *The following relationships hold.*

(a) $\|\mathbf{W}\mathbf{x} - \mathbf{J}\mathbf{x}\| \leq \lambda \|\mathbf{x} - \mathbf{J}\mathbf{x}\|, \forall \mathbf{x} \in \mathbb{R}^{np}$.

(b) $\overline{\mathbf{y}}_{t+1} = \overline{\mathbf{v}}_t, \forall t \geq 0$.

(c) $\left\|\overline{\nabla \mathbf{f}}(\mathbf{x}_t) - \nabla F(\overline{\mathbf{x}}_t)\right\|^2 \leq \frac{L^2}{n} \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2, \forall t \geq 0$.

We note that Lemma 5.5.1(a) holds since $\underline{\mathbf{W}}$ is primitive and doubly-stochastic, Lemma 5.5.1(b) is a direct consequence of the gradient tracking update (5.4a) and Lemma 5.5.1(c) is due to the L-smoothness of each f_i. By the update of `GT-HSGD` described in (5.4b) and Lemma 5.5.1(b), it is straightforward to obtain:

$$\overline{\mathbf{x}}_{t+1} = \overline{\mathbf{x}}_t - \alpha \overline{\mathbf{y}}_{t+1} = \overline{\mathbf{x}}_t - \alpha \overline{\mathbf{v}}_t, \qquad \forall t \geq 0. \tag{5.5}$$

Hence, the mean state $\overline{\mathbf{x}}_t$ proceeds in the direction of the average of local stochastic gradient estimators $\overline{\mathbf{v}}_t$. With the help of (5.5) and the L-smoothness of F and each f_i, we establish the following descent inequality which sheds light on the overall convergence analysis.

Lemma 5.5.2. *If* $0 < \alpha \leq \frac{1}{2L}$, *then we have:* $\forall T \geq 0$,

$$\sum_{t=0}^{T} \left\|\nabla F(\overline{\mathbf{x}}_t)\right\|^2 \leq \frac{2(F(\overline{\mathbf{x}}_0) - F^*)}{\alpha} - \frac{1}{2}\sum_{t=0}^{T} \|\overline{\mathbf{v}}_t\|^2 + 2\sum_{t=0}^{T} \left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + \frac{2L^2}{n}\sum_{t=0}^{T} \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2.$$

Proof. See Section 5.6.1. □

In light of Lemma 5.5.2, our approach to establishing the convergence of `GT-HSGD` is to seek the conditions on the algorithmic parameters of `GT-HSGD`, i.e., the step-size α and the weight parameter β, such that

$$-\frac{1}{2T}\sum_{t=0}^{T} \mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{2}{T}\sum_{t=0}^{T} \mathbb{E}\left[\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\|^2\right] + \frac{2L^2}{nT}\sum_{t=0}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] = O\left(\alpha, \beta, \frac{1}{b_0}, \frac{1}{T}\right), \tag{5.6}$$

where $O(\alpha, \beta, 1/b_0, 1/T)$ represents a nonnegative quantity which may be made arbitrarily small by choosing small enough α and β along with large enough T and b_0. If (5.6) holds, then Lemma 5.5.2 reduces to

$$\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}\left[\|\nabla F(\overline{\mathbf{x}}_t)\|^2\right] \leq \frac{2(F(\overline{\mathbf{x}}_0) - F^*)}{\alpha T} + O\left(\alpha, \beta, \frac{1}{b_0}, \frac{1}{T}\right),$$

which leads to the convergence of `GT-HSGD`. To this aim, we quantify $\mathbb{E}\big[\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\|^2\big]$ and $\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\big]$.

5.5.1 Contraction relationships

First of all, we bound the gradient variances by exploiting the hybrid and recursive update of $\mathbf{v}_t$.

Lemma 5.5.3. *The following inequalities hold:* $\forall t \geq 1$,

$$\mathbb{E}\left[\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\|^2\right] \leq (1-\beta)^2 \mathbb{E}\left[\|\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\|^2\right] + \frac{6L^2\alpha^2}{n}(1-\beta)^2 \mathbb{E}\left[\|\overline{\mathbf{v}}_{t-1}\|^2\right] + \frac{2\beta^2\overline{\nu}^2}{n} + \frac{6L^2}{n^2}(1-\beta)^2 \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + \|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2\right], \tag{5.7}$$

$$\mathbb{E}\left[\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{x}_t)\|^2\right] \leq (1-\beta)^2 \mathbb{E}\left[\|\mathbf{v}_{t-1} - \nabla \mathbf{f}(\mathbf{x}_{t-1})\|^2\right] + 6nL^2\alpha^2(1-\beta)^2 \mathbb{E}\left[\|\overline{\mathbf{v}}_{t-1}\|^2\right] + 2n\beta^2\overline{\nu}^2 + 6L^2(1-\beta)^2 \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + \|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2\right]. \tag{5.8}$$

Proof. See Section 5.6.2. □

Remark 5.5.1. Since $\mathbf{v}_t$ is a conditionally biased estimator of $\nabla \mathbf{f}(\mathbf{x}_t)$, (5.7) and (5.8) do not directly imply each other and need to be established separately.

We emphasize that the contraction structure of the gradient variances shown in Lemma 5.5.3 plays a crucial role in the convergence analysis. The following contraction bounds on the consensus errors $\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right]$ are standard in decentralized algorithms based on gradient tracking, e.g., [67, 141]; in particular, it follows directly from the $\mathbf{x}_t$-update (5.4b) and Young's inequality.

Lemma 5.5.4. *The following inequalities hold:* $\forall t \geq 0$,

$$\|\mathbf{x}_{t+1} - \mathbf{J}\mathbf{x}_{t+1}\|^2 \leq \frac{1+\lambda^2}{2}\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + \frac{2\alpha^2\lambda^2}{1-\lambda^2}\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2. \tag{5.9}$$

$$\|\mathbf{x}_{t+1} - \mathbf{J}\mathbf{x}_{t+1}\|^2 \leq 2\lambda^2\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + 2\alpha^2\lambda^2\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2. \tag{5.10}$$

It is then clear from Lemma 5.5.4 that we need to further quantify the gradient tracking errors $\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right]$ in order to bound the consensus errors. These error bounds are shown in the following lemma.

Lemma 5.5.5. *We have the following.*

(a) $\mathbb{E}\left[\|\mathbf{y}_1 - \mathbf{J}\mathbf{y}_1\|^2\right] \leq \lambda^2\left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2 + \lambda^2 n\overline{\nu}^2/b_0$.

(b) If $0 < \alpha \leq \frac{1-\lambda^2}{2\sqrt{42\lambda^2}L}$, *then* $\forall t \geq 1$,

$$\mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2\right] \leq \frac{3+\lambda^2}{4}\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right] + \frac{21\lambda^2 nL^2\alpha^2}{1-\lambda^2}\mathbb{E}[\|\overline{\mathbf{v}}_{t-1}\|^2] + \frac{63\lambda^2L^2}{1-\lambda^2}\mathbb{E}\left[\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2\right] + \frac{7\lambda^2\beta^2}{1-\lambda^2}\mathbb{E}\left[\|\mathbf{v}_{t-1} - \nabla \mathbf{f}(\mathbf{x}_{t-1})\|^2\right] + 3\lambda^2 n\beta^2\overline{\nu}^2.$$

Proof. See Section 5.6.3. □

We note that establishing the contraction argument of gradient tracking errors in Lemma 5.5.5 requires a careful examination of the structure of the $\mathbf{v}_t$-update.

5.5.2 Error accumulations

To proceed, we observe, from Lemma 5.5.3, 5.5.4, and 5.5.5, that the recursions of the gradient variances, consensus, and gradient tracking errors admit similar forms. Therefore, we abstract out formulas for the accumulation of the error recursions of this type in the following lemma.

Lemma 5.5.6. *Let $\{V_t\}_{t\geq 0}$, $\{R_t\}_{t\geq 0}$ and $\{Q_t\}_{t\geq 0}$ be nonnegative sequences and $C \geq 0$ be some constant such that $V_t \leq qV_{t-1}+qR_{t-1}+Q_t+C$, $\forall t \geq 1$, where $q \in (0,1)$. Then the following inequality holds: $\forall T \geq 1$,*

$$\sum_{t=0}^{T} V_t \leq \frac{V_0}{1-q} + \frac{1}{1-q}\sum_{t=0}^{T-1} R_t + \frac{1}{1-q}\sum_{t=1}^{T} Q_t + \frac{CT}{1-q}. \tag{5.11}$$

Similarly, if $V_{t+1} \leq qV_t + R_{t-1} + C, \forall t \geq 1$, then we have: $\forall T \geq 2$,

$$\sum_{t=1}^{T} V_t \leq \frac{V_1}{1-q} + \frac{1}{1-q}\sum_{t=0}^{T-2} R_t + \frac{CT}{1-q}. \tag{5.12}$$

Proof. See Section 5.6.4. □

Applying Lemma 5.5.6 to Lemma 5.5.3 leads to the following upper bounds on the accumulated variances.

Lemma 5.5.7. *For any $\beta \in (0,1)$, the following inequalities hold: $\forall T \geq 1$,*

$$\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t - \overline{\nabla\mathbf{f}}(\mathbf{x}_t)\|^2\right] \leq \frac{\overline{\nu}^2}{\beta b_0 n} + \frac{6L^2\alpha^2}{n\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{12L^2}{n^2\beta}\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] + \frac{2\beta\overline{\nu}^2 T}{n}, \tag{5.13}$$

$$\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{x}_t)\|^2\right] \leq \frac{n\overline{\nu}^2}{\beta b_0} + \frac{6nL^2\alpha^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{12L^2}{\beta}\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] + 2n\beta\overline{\nu}^2 T. \tag{5.14}$$

Proof. See Section 5.6.5. □

It can be observed that (5.13) in Lemma 5.5.7 may be used to refine the left hand side of (5.6). The remaining step, naturally, is to bound $\sum_t \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right]$ in terms of $\sum_t \mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right]$. This result is provided in the following lemma that is obtained with the help of Lemma 5.5.4, 5.5.5, 5.5.6, and 5.5.7.

Lemma 5.5.8. *If $0 < \alpha \leq \frac{(1-\lambda^2)^2}{70\lambda^2 L}$ and $\beta \in (0,1)$, then the following inequality holds: $\forall T \geq 2$,*

$$\begin{aligned}\sum_{t=0}^{T}\frac{\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right]}{n} \leq{}& \frac{2016\lambda^4 L^2\alpha^4}{(1-\lambda^2)^4}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{32\lambda^4\alpha^2}{(1-\lambda^2)^3}\frac{\|\nabla\mathbf{f}(\mathbf{x}_0)\|^2}{n} + \left(\frac{7\beta}{1-\lambda^2}+1\right)\frac{32\lambda^4\overline{\nu}^2\alpha^2}{(1-\lambda^2)^3 b_0}\\ &+ \left(\frac{14\beta}{1-\lambda^2}+3\right)\frac{32\lambda^4\beta^2\overline{\nu}^2\alpha^2 T}{(1-\lambda^2)^3}.\end{aligned}$$

Proof. See Section 5.6.6. □

Finally, we note that Lemma 5.5.7 and 5.5.8 suffice to establish (5.6) and hence lead to Theorem 5.3.1, whose detailed proof is presented in the next subsection.

5.5.3 Proof of Theorem 5.3.1

For the ease of presentation, we denote $\Delta_0 := F(\overline{\mathbf{x}}_0) - F^*$ in the following. We apply (5.13) to Lemma 5.5.2 to obtain: if $0 < \alpha \leq \frac{1}{2L}$, then $\forall T \geq 1$,

$$\begin{aligned}\sum_{t=0}^{T} \mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}_t)\right\|^2\right] &\leq \frac{2\Delta_0}{\alpha} - \frac{1}{2}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{2L^2}{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] \\ &\quad + \frac{2\overline{\nu}^2}{\beta b_0 n} + \frac{12L^2\alpha^2}{n\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{24L^2}{n^2\beta}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + \frac{4\beta\overline{\nu}^2 T}{n} \\ &\leq \frac{2\Delta_0}{\alpha} - \frac{1}{4}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{2L^2}{n}\left(1 + \frac{12}{n\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] \\ &\quad + \frac{2\overline{\nu}^2}{\beta b_0 n} + \frac{4\beta\overline{\nu}^2 T}{n} - \left(\frac{1}{4} - \frac{12L^2\alpha^2}{n\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right]. \end{aligned} \tag{5.15}$$

Therefore, if $0 < \alpha < \frac{1}{4\sqrt{3}L}$ and $\frac{48L^2\alpha^2}{n} \leq \beta < 1$, i.e., $\frac{1}{4} - \frac{12L^2\alpha^2}{n\beta} \geq 0$, we may drop the last term in (5.15) to obtain: $\forall T \geq 1$,

$$\sum_{t=0}^{T} \mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}_t)\right\|^2\right] \leq \frac{2\Delta_0}{\alpha} - \frac{1}{4}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{2L^2}{n}\left(1 + \frac{12}{n\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + \frac{2\overline{\nu}^2}{\beta b_0 n} + \frac{4\beta\overline{\nu}^2 T}{n}. \tag{5.16}$$

Moreover, we observe: $\forall T \geq 1$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i)\right\|^2\right] &\leq \frac{2}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i) - \nabla F(\overline{\mathbf{x}}_t)\right\|^2 + \left\|\nabla F(\overline{\mathbf{x}}_t)\right\|^2\right] \\ &= \frac{2L^2}{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + 2\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\overline{\mathbf{x}}_t)\right\|^2\right], \end{aligned} \tag{5.17}$$

where the last line uses the L-smoothness of F. Using (5.16) in (5.17) yields: if $0 < \alpha < \frac{1}{4\sqrt{3}L}$ and $48L^2\alpha^2/n \leq \beta < 1$, then we have: $\forall T \geq 1$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i)\right\|^2\right] &\leq \frac{4\Delta_0}{\alpha} - \frac{1}{2}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{6L^2}{n}\left(1 + \frac{8}{n\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] \\ &\quad + \frac{4\overline{\nu}^2}{\beta b_0 n} + \frac{8\beta\overline{\nu}^2 T}{n}. \end{aligned} \tag{5.18}$$

According to (5.18), if $0 < \alpha < \frac{1}{4\sqrt{3}L}$ and $\beta = 48L^2\alpha^2/n$, we have: $\forall T \geq 1$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i)\right\|^2\right] &\leq \frac{4\Delta_0}{\alpha} - \frac{1}{2}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{6L^2}{n}\left(1 + \frac{1}{6L^2\alpha^2}\right)\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] \\ &\quad + \frac{4\overline{\nu}^2}{\beta b_0 n} + \frac{8\beta\overline{\nu}^2 T}{n} \\ &\leq \frac{4\Delta_0}{\alpha} \underbrace{- \frac{1}{2}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{2}{n\alpha^2}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right]}_{=:\Phi_T} + \frac{4\overline{\nu}^2}{\beta b_0 n} + \frac{8\beta\overline{\nu}^2 T}{n}, \end{aligned} \tag{5.19}$$

where the last line is due to $6L^2\alpha^2 < 1/8$. To simplify Φ_T, we use Lemma 5.5.8 to obtain: if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{70\lambda^2 L}$ then $\forall T \geq 2$,

$$\Phi_T \leq -\frac{1}{2}\left(1 - \frac{8064\lambda^4 L^2\alpha^2}{(1-\lambda^2)^4}\right)\sum_{t=0}^{T}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right] + \frac{64\lambda^4}{(1-\lambda^2)^3}\frac{\left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2}{n} + \left(\frac{7\beta}{1-\lambda^2}+1\right)\frac{64\lambda^4\overline{\nu}^2}{(1-\lambda^2)^3 b_0} + \left(\frac{14\beta}{1-\lambda^2}+3\right)\frac{64\lambda^4\beta^2\overline{\nu}^2 T}{(1-\lambda^2)^3}. \tag{5.20}$$

In (5.20), we observe that if $0 < \alpha \leq \frac{(1-\lambda^2)^2}{90\lambda^2 L}$, then $1 - \frac{8064\lambda^4 L^2\alpha^2}{(1-\lambda^2)^4} \geq 0$ and thus the first term in (5.20) may be dropped; moreover, if $0 < \alpha \leq \frac{\sqrt{n(1-\lambda^2)}}{26\lambda L}$, then $\beta = \frac{48L^2\alpha^2}{n} \leq \frac{1-\lambda^2}{14\lambda^2}$. Hence, if $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{90\lambda^2}, \frac{\sqrt{n(1-\lambda^2)}}{26\lambda}\right\}\frac{1}{L}$, then (5.20) reduces to: $\forall T \geq 2$,

$$\Phi_T \leq \frac{64\lambda^4}{(1-\lambda^2)^3}\frac{\left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2}{n} + \frac{96\lambda^2\overline{\nu}^2}{(1-\lambda^2)^3 b_0} + \frac{256\lambda^2\beta^2\overline{\nu}^2 T}{(1-\lambda^2)^3}. \tag{5.21}$$

Finally, we use (5.21) in (5.19) to obtain: if $0 < \alpha < \min\left\{\frac{1}{4\sqrt{3}}, \frac{(1-\lambda^2)^2}{90\lambda^2}, \frac{\sqrt{n(1-\lambda^2)}}{26\lambda}\right\}\frac{1}{L}$, we have: $\forall T \geq 2$,

$$\frac{1}{n(T+1)}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t^i)\right\|^2\right] \leq \frac{4\Delta_0}{\alpha T} + \frac{4\overline{\nu}^2}{\beta b_0 n T} + \frac{8\beta\overline{\nu}^2}{n} + \frac{64\lambda^4}{(1-\lambda^2)^3 T}\frac{\left\|\nabla \mathbf{f}(\mathbf{x}_0)\right\|^2}{n} + \frac{96\lambda^2\overline{\nu}^2}{(1-\lambda^2)^3 b_0 T} + \frac{256\lambda^2\beta^2\overline{\nu}^2}{(1-\lambda^2)^3}. \tag{5.22}$$

The proof follows by (5.22) and that $\mathbb{E}[\|\nabla F(\widetilde{\mathbf{x}}_T)\|^2] = \frac{1}{n(T+1)}\sum_{i=1}^{n}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t^i)\|^2]$ since $\widetilde{\mathbf{x}}_T$ is chosen uniformly at random from $\{\mathbf{x}_t^i : \forall i \in \mathcal{V}, 0 \leq t \leq T\}$.

5.6 Detailed proofs for lemmata in Section 5.5

5.6.1 Proof of Lemma 5.5.2

We recall the standard Descent Lemma [6], i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2, \tag{5.23}$$

since F is L-smooth. Setting $\mathbf{y} = \overline{\mathbf{x}}_{t+1}$ and $\mathbf{x} = \overline{\mathbf{x}}_t$ in (5.23) and using (5.5) , we have: $\forall t \geq 0$,

$$\begin{aligned} F(\overline{\mathbf{x}}_{t+1}) &\leq F(\overline{\mathbf{x}}_t) - \left\langle \nabla F(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_{t+1} - \overline{\mathbf{x}}_t\right\rangle + \frac{L}{2}\|\overline{\mathbf{x}}_{t+1} - \overline{\mathbf{x}}_t\|^2 \\ &\leq F(\overline{\mathbf{x}}_t) - \alpha\left\langle \nabla F(\overline{\mathbf{x}}_t), \overline{\mathbf{v}}_t\right\rangle + \frac{L\alpha^2}{2}\|\overline{\mathbf{v}}_t\|^2. \end{aligned} \tag{5.24}$$

Using $\langle \mathbf{a}, \mathbf{b}\rangle = \frac{1}{2}\left(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2\right), \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, in (5.24) gives: for $0 < \alpha \leq \frac{1}{2L}$ and $\forall t \geq 0$,

$$\begin{aligned} F(\overline{\mathbf{x}}_{t+1}) &\leq F(\overline{\mathbf{x}}_t) - \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}_t)\|^2 - \left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right)\|\overline{\mathbf{v}}_t\|^2 + \frac{\alpha}{2}\|\overline{\mathbf{v}}_t - \nabla F(\overline{\mathbf{x}}_t)\|^2, \\ &\leq F(\overline{\mathbf{x}}_t) - \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}_t)\|^2 - \left(\frac{\alpha}{2} - \frac{L\alpha^2}{2}\right)\|\overline{\mathbf{v}}_t\|^2 + \alpha\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + \alpha\left\|\overline{\nabla \mathbf{f}}(\mathbf{x}_t) - \nabla F(\overline{\mathbf{x}}_t)\right\|^2, \\ &\overset{(i)}{\leq} F(\overline{\mathbf{x}}_t) - \frac{\alpha}{2}\|\nabla F(\overline{\mathbf{x}}_t)\|^2 - \frac{\alpha}{4}\|\overline{\mathbf{v}}_t\|^2 + \alpha\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + \frac{\alpha L^2}{n}\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2, \end{aligned} \tag{5.25}$$

where (i) is due to Lemma 5.5.1(c) and that $\frac{L\alpha^2}{2} \leq \frac{\alpha}{4}$ since $0 < \alpha \leq \frac{1}{2L}$. Rearranging (5.25), we have: for $0 < \alpha \leq \frac{1}{2L}$ and $\forall t \geq 0$,

$$\|\nabla F(\overline{\mathbf{x}}_t)\|^2 \leq \frac{2(F(\overline{\mathbf{x}}_t) - F(\overline{\mathbf{x}}_{t+1}))}{\alpha} - \frac{1}{2}\|\overline{\mathbf{v}}_t\|^2 + 2\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + \frac{2L^2}{n}\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2. \tag{5.26}$$

Taking the telescoping sum of (5.26) over t from 0 to T, $\forall T \geq 0$ and using the fact that F bounded below by F^* in the resulting inequality finishes the proof.

5.6.2 Proof of Lemma 5.5.3

5.6.2.1 Proof of Eq. (5.7)

We recall that the update of each local stochastic gradient estimator $\mathbf{v}_t^i, \forall t \geq 1$, in (5.2) may be written equivalently as follows:

$$\mathbf{v}_t^i = \beta \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) + (1-\beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \mathbf{v}_{t-1}^i\Big),$$

where $\beta \in (0,1)$. We have: $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\begin{aligned}
\mathbf{v}_t^i - \nabla f_i(\mathbf{x}_t^i) &= \beta \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) + (1-\beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \mathbf{v}_{t-1}^i\Big) - \beta\nabla f_i(\mathbf{x}_t^i) - (1-\beta)\nabla f_i(\mathbf{x}_t^i) \\
&= \beta\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\Big) + (1-\beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_t^i)\Big) \\
&= \beta\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\Big) + (1-\beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) - \nabla f_i(\mathbf{x}_t^i)\Big) \\
&\quad + (1-\beta)\Big(\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\Big).
\end{aligned} \tag{5.27}$$

In (5.27), we observe that $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\mathbb{E}\Big[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i) | \mathcal{F}_t\Big] = \mathbf{0}_p, \tag{5.28}$$

$$\mathbb{E}\Big[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) - \nabla f_i(\mathbf{x}_t^i) | \mathcal{F}_t\Big] = \mathbf{0}_p, \tag{5.29}$$

by the definition of the filtration $\mathcal{F}_t$ in (5.1). Averaging (5.27) over i from 1 to n gives: $\forall t \geq 0$,

$$\begin{aligned}
\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t) &= (1-\beta)\Big(\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\Big) \\
&\quad + \beta \cdot \underbrace{\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\Big)}_{=:\mathbf{s}_t} \\
&\quad + (1-\beta) \cdot \underbrace{\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) - \nabla f_i(\mathbf{x}_t^i)\Big)}_{=:\mathbf{z}_t}.
\end{aligned} \tag{5.30}$$

Note that $\mathbb{E}[\mathbf{s}_t|\mathcal{F}_t] = \mathbb{E}[\mathbf{z}_t|\mathcal{F}_t] = \mathbf{0}_p$ by (5.28) and (5.29). In light of (5.30), we have: $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\Big[\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\|^2|\mathcal{F}_t\Big] &= (1-\beta)^2\|\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\|^2 + \mathbb{E}\left[\|\beta\sigma_t + (1-\beta)\mathbf{z}_t\|^2|\mathcal{F}_t\right] \\
&\quad + 2\mathbb{E}\Big[\Big\langle(1-\beta)\left(\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\right), \beta\sigma_t + (1-\beta)\mathbf{z}_t\Big\rangle|\mathcal{F}_t\Big] \\
&\overset{(i)}{=} (1-\beta)^2\|\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\|^2 + \mathbb{E}\left[\|\beta\sigma_t + (1-\beta)\mathbf{z}_t\|^2|\mathcal{F}_t\right] \\
&\leq (1-\beta)^2\|\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\|^2 + 2\beta^2\mathbb{E}\left[\|\sigma_t\|^2|\mathcal{F}_t\right] + 2(1-\beta)^2\mathbb{E}\left[\|\mathbf{z}_t\|^2|\mathcal{F}_t\right], \quad (5.31)
\end{aligned}$$

where (i) is due to

$$\mathbb{E}\Big[\Big\langle(1-\beta)\left(\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1})\right), \beta\sigma_t + (1-\beta)\mathbf{z}_t\Big\rangle|\mathcal{F}_t\Big] = 0,$$

since $\mathbb{E}[\sigma_t|\mathcal{F}_t] = \mathbb{E}[\mathbf{z}_t|\mathcal{F}_t] = \mathbf{0}_p$ and $(\overline{\mathbf{v}}_{t-1} - \overline{\nabla \mathbf{f}}(\mathbf{x}_{t-1}))$ is $\mathcal{F}_t$-measurable. We next bound the second and the third term in (5.31) respectively. For the second term in (5.31), we observe that $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\|\sigma_t\|^2\right] &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\right\|^2\right] + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\left[\left\langle \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i), \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \nabla f_j(\mathbf{x}_t^j)\right\rangle\right] \\
&\overset{(i)}{=} \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\right\|^2\right] \leq \frac{\overline{\nu}^2}{n}. \quad (5.32)
\end{aligned}$$

We note that (i) in (5.32) uses that whenever $i \neq j$,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i), \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \nabla f_j(\mathbf{x}_t^j)\right\rangle|\mathcal{F}_t\right] \\
&\overset{(ii)}{=} \mathbb{E}\left[\left\langle \mathbb{E}\left[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i)|\sigma(\boldsymbol{\xi}_t^j, \mathcal{F}_t)\right] - \nabla f_i(\mathbf{x}_t^i), \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \nabla f_j(\mathbf{x}_t^j)\right\rangle|\mathcal{F}_t\right] \\
&\overset{(iii)}{=} \mathbb{E}\left[\left\langle \mathbb{E}\left[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i)|\mathcal{F}_t\right] - \nabla f_i(\mathbf{x}_t^i), \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \nabla f_j(\mathbf{x}_t^j)\right\rangle\Big|\mathcal{F}_t\right] = 0, \quad (5.33)
\end{aligned}$$

where (ii) is due to the tower property of the conditional expectation and (iii) uses that $\boldsymbol{\xi}_t^j$ is independent of $\{\boldsymbol{\xi}_t^i, \mathcal{F}_t\}$ and $\mathbf{x}_t^i$ is $\mathcal{F}_t$-measurable. Towards the third term (5.31), we define, $\forall t \geq 1$,

$$\widehat{\nabla}_t^i := \nabla f_i(\mathbf{x}_t^i) - \nabla f_i(\mathbf{x}_{t-1}^i)$$

and recall that $\mathbb{E}\left[\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)|\mathcal{F}_t\right] = \widehat{\nabla}_t^i$. Observe that $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{z}_t\|^2|\mathcal{F}_t\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \widehat{\nabla}_t^i\right)\right\|^2\Big|\mathcal{F}_t\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \widehat{\nabla}_t^i\right\|^2|\mathcal{F}_t\right] \\
&\quad + \frac{1}{n^2}\sum_{i\neq j}\underbrace{\mathbb{E}\left[\left\langle \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \widehat{\nabla}_t^i, \mathbf{g}_j(\mathbf{x}_t^j, \boldsymbol{\xi}_t^j) - \mathbf{g}_j(\mathbf{x}_{t-1}^j, \boldsymbol{\xi}_t^j) - \widehat{\nabla}_t^j\right\rangle|\mathcal{F}_t\right]}_{=0} \\
&\overset{(i)}{=} \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \widehat{\nabla}_t\right\|^2|\mathcal{F}_t\right], \\
&\overset{(ii)}{\leq} \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)\right\|^2|\mathcal{F}_t\right], \quad (5.34)
\end{aligned}$$

where (i) follows from a similar line of arguments as (5.33) and (ii) uses the conditional variance decomposition, i.e., for any random vector $\mathbf{a} \in \mathbb{R}^p$ consisted of square-integrable random variables,

$$\mathbb{E}\left[\left\|\mathbf{a} - \mathbb{E}\left[\mathbf{a}|\mathcal{F}_t\right]\right\|^2 |\mathcal{F}_t\right] = \mathbb{E}\left[\|\mathbf{a}\|^2 |\mathcal{F}_t\right] - \left\|\mathbb{E}\left[\mathbf{a}|\mathcal{F}_t\right]\right\|^2. \tag{5.35}$$

To proceed from (5.34), we take its expectation and observe that $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{z}_t\|^2\right] &\leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)\right\|^2\right] \\
&\overset{(i)}{\leq} \frac{L^2}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{x}_t^i - \mathbf{x}_{t-1}^i\right\|^2\right] \\
&= \frac{L^2}{n^2}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}_{t-1}\right\|^2\right] \\
&= \frac{L^2}{n^2}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t + \mathbf{J}\mathbf{x}_t - \mathbf{J}\mathbf{x}_{t-1} + \mathbf{J}\mathbf{x}_{t-1} - \mathbf{x}_{t-1}\right\|^2\right] \\
&\leq \frac{3L^2}{n^2}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + n\left\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_{t-1}\right\|^2 + \left\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\right\|^2\right] \\
&\overset{(ii)}{=} \frac{3L^2\alpha^2}{n}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_{t-1}\right\|^2\right] + \frac{3L^2}{n^2}\Big(\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \left\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\right\|^2\right]\Big),
\end{aligned} \tag{5.36}$$

where (i) uses the mean-squared smoothness of each $\mathbf{g}_i$ and (ii) uses the update of $\overline{\mathbf{x}}_t$ in (5.5). The proof follows by taking the expectation (5.31) and then using (5.32) and (5.36) in the resulting inequality.

5.6.2.2 Proof of Eq. (5.8)

We recall from (5.27) the following relationship: $\forall t \geq 1$,

$$\begin{aligned}
\mathbf{v}_t^i - \nabla f_i(\mathbf{x}_t^i) = \beta\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\Big) + (1-\beta)\Big(\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) - \nabla f_i(\mathbf{x}_t^i)\Big) \\
+ (1-\beta)\Big(\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\Big).
\end{aligned} \tag{5.37}$$

Note that the conditional expectation of the first and second term in (5.37) given $\mathcal{F}_t$ is 0 and that the third term in (5.37) is $\mathcal{F}_t$-measurable. Following a similar procedure in the proof of (5.31), we have: $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{v}_t^i - \nabla f_i(\mathbf{x}_t^i)\|^2 |\mathcal{F}_t\right] &\leq (1-\beta)^2\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2 + 2\beta^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\right\|^2 |\mathcal{F}_t\right] \\
&\quad + 2(1-\beta)^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \big(\nabla f_i(\mathbf{x}_t^i) - \nabla f_i(\mathbf{x}_{t-1}^i)\big)\right\|^2 |\mathcal{F}_t\right] \\
&\overset{(i)}{\leq} (1-\beta)^2\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2 + 2\beta^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_t^i)\right\|^2 |\mathcal{F}_t\right] \\
&\quad + 2(1-\beta)^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)\right\|^2 |\mathcal{F}_t\right]
\end{aligned} \tag{5.38}$$

where (i) uses the conditional variance decomposition (5.35). We then take the expectation of (5.38) with the help of the mean-squared smoothness and the bounded variance of each $\mathbf{g}_i$ to proceed: $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{v}_t^i - \nabla f_i(\mathbf{x}_t^i)\right\|^2\right] &\leq (1-\beta)^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] + 2\beta^2\nu_i^2 + 2(1-\beta)^2L^2\mathbb{E}\left[\left\|\mathbf{x}_t^i - \mathbf{x}_{t-1}^i\right\|^2\right] \\
&\leq (1-\beta)^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] + 2\beta^2\nu_i^2 \\
&\quad + 6(1-\beta)^2L^2\left(\mathbb{E}\left[\left\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\right\|^2 + \left\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_{t-1}\right\|^2 + \left\|\overline{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}^i\right\|^2\right]\right), \\
&= (1-\beta)^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] + 2\beta^2\nu_i^2 + 6(1-\beta)^2L^2\alpha^2\mathbb{E}\left[\left\|\overline{\mathbf{v}}_{t-1}\right\|^2\right] \\
&\quad + 6(1-\beta)^2L^2\mathbb{E}\left[\left\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\right\|^2 + \left\|\mathbf{x}_{t-1}^i - \overline{\mathbf{x}}_{t-1}\right\|^2\right], \qquad (5.39)
\end{aligned}$$

where the last line uses the $\overline{\mathbf{x}}_t$-update in (5.5). Summing up (5.39) over i from 1 to n completes the proof.

5.6.3 Proof of Lemma 5.5.5

5.6.3.1 Proof of Lemma 5.5.5(a)

Recall the initialization of `GT-HSGD` that $\mathbf{v}_{-1} = \mathbf{0}_{np}$, $\mathbf{y}_0 = \mathbf{0}_{np}$, and $\mathbf{v}_0^i = \frac{1}{b_0}\sum_{r=1}^{b_0}\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i)$. Using the gradient tracking update (5.4a) at iteration $t = 0$, we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{y}_1 - \mathbf{J}\mathbf{y}_1\right\|^2\right] &= \mathbb{E}\left[\left\|\mathbf{W}\left(\mathbf{y}_0 + \mathbf{v}_0 - \mathbf{v}_{-1}\right) - \mathbf{J}\mathbf{W}\left(\mathbf{y}_0 + \mathbf{v}_0 - \mathbf{v}_{-1}\right)\right\|^2\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\left\|\left(\mathbf{W} - \mathbf{J}\right)\mathbf{v}_0\right\|^2\right] \\
&\overset{(ii)}{\leq} \lambda^2\mathbb{E}\left[\left\|\mathbf{v}_0 - \nabla\mathbf{f}(\mathbf{x}_0) + \nabla\mathbf{f}(\mathbf{x}_0)\right\|^2\right] \\
&= \lambda^2\sum_{i=1}^n\mathbb{E}\left[\left\|\mathbf{v}_0^i - \nabla f_i(\mathbf{x}_0^i)\right\|^2\right] + \lambda^2\left\|\nabla\mathbf{f}(\mathbf{x}_0)\right\|^2 \\
&\overset{(iii)}{=} \lambda^2\sum_{i=1}^n\mathbb{E}\left[\left\|\frac{1}{b_0}\sum_{r=1}^{b_0}\left(\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right)\right\|^2\right] + \lambda^2\left\|\nabla\mathbf{f}(\mathbf{x}_0)\right\|^2 \\
&\overset{(iv)}{=} \frac{\lambda^2}{b_0^2}\sum_{i=1}^n\sum_{r=1}^{b_0}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right\|^2\right] + \lambda^2\left\|\nabla\mathbf{f}(\mathbf{x}_0)\right\|^2, \qquad (5.40)
\end{aligned}$$

where (i) uses $\mathbf{J}\mathbf{W} = \mathbf{J}$ and the initial condition of $\mathbf{v}_{-1}$ and $\mathbf{y}_0$, (ii) uses $\|\mathbf{W} - \mathbf{J}\| = \lambda$, (iii) is due to the initialization of $\mathbf{v}_0^i$, and (iv) follows from the fact that $\{\boldsymbol{\xi}_{0,1}^i, \boldsymbol{\xi}_{0,2}^i, \cdots, \boldsymbol{\xi}_{0,b_0}^i\}$, $\forall i \in \mathcal{V}$, is an independent family of random vectors, by a similar line of arguments in (5.32) and (5.33). The proof then follows by using the bounded variance of each $\mathbf{g}_i$ in (5.40).

5.6.3.2 Proof of Lemma 5.5.5(b)

Following the gradient tracking update (5.4a), we have: $\forall t \geq 1$,

$$\begin{aligned}
\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 =& \|\mathbf{W}(\mathbf{y}_t + \mathbf{v}_t - \mathbf{v}_{t-1}) - \mathbf{J}\mathbf{W}(\mathbf{y}_t + \mathbf{v}_t - \mathbf{v}_{t-1})\|^2 \\
\overset{(i)}{=}& \|\mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t + (\mathbf{W} - \mathbf{J})(\mathbf{v}_t - \mathbf{v}_{t-1})\|^2 \\
=& \|\mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + 2\big\langle \mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, (\mathbf{W} - \mathbf{J})(\mathbf{v}_t - \mathbf{v}_{t-1})\big\rangle + \|(\mathbf{W} - \mathbf{J})(\mathbf{v}_t - \mathbf{v}_{t-1})\|^2 \\
\overset{(ii)}{\leq}& \lambda^2 \|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \underbrace{2\big\langle \mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, (\mathbf{W} - \mathbf{J})(\mathbf{v}_t - \mathbf{v}_{t-1})\big\rangle}_{=:A_t} + \lambda^2 \|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2, \qquad (5.41)
\end{aligned}$$

where (i) uses $\mathbf{J}\mathbf{W} = \mathbf{J}$ and (ii) is due to $\|\mathbf{W} - \mathbf{J}\| = \lambda$. In the following, we bound A_t and the last term in (5.41) respectively. We recall the update of each local stochastic gradient estimator $\mathbf{v}_t^i$ in (5.2): $\forall t \geq 1$,

$$\mathbf{v}_t^i = \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) + (1-\beta)\mathbf{v}_{t-1}^i - (1-\beta)\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i).$$

We observe that $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\begin{aligned}
\mathbf{v}_t^i - \mathbf{v}_{t-1}^i =& \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \beta\mathbf{v}_{t-1}^i - (1-\beta)\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) \\
=& \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \beta\mathbf{v}_{t-1}^i + \beta\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) \\
=& \mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \beta\Big(\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\Big) + \beta\Big(\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_{t-1}^i)\Big). \qquad (5.42)
\end{aligned}$$

Moreover, we observe from (5.42) that $\forall t \geq 1$,

$$\mathbb{E}[\mathbf{v}_t - \mathbf{v}_{t-1}|\mathcal{F}_t] = \nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1}) - \beta\Big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\Big). \qquad (5.43)$$

Towards A_t, we have: $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}[A_t|\mathcal{F}_t] \overset{(i)}{=}& 2\Big\langle \mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, (\mathbf{W} - \mathbf{J})\,\mathbb{E}[\mathbf{v}_t - \mathbf{v}_{t-1}|\mathcal{F}_t]\Big\rangle \\
\overset{(ii)}{=}& 2\Big\langle \mathbf{W}\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, (\mathbf{W} - \mathbf{J})\Big(\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1}) - \beta\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)\Big)\Big\rangle \\
\overset{(iii)}{\leq}& 2\lambda\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\| \cdot \lambda\Big\|\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1}) - \beta\Big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\Big)\Big\| \\
\overset{(iv)}{\leq}& \frac{1-\lambda^2}{2}\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \frac{2\lambda^4}{1-\lambda^2}\Big\|\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1}) - \beta\Big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\Big)\Big\|^2, \\
\overset{(v)}{\leq}& \frac{1-\lambda^2}{2}\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \frac{4\lambda^4 L^2}{1-\lambda^2}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \frac{4\lambda^4\beta^2}{1-\lambda^2}\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\|^2, \qquad (5.44)
\end{aligned}$$

where (i) is due to the $\mathcal{F}_t$-measurability of $\mathbf{y}_t$, (ii) uses (5.43), (iii) is due to the Cauchy-Schwarz inequality and $\|\mathbf{W} - \mathbf{J}\| = \lambda$, (iv) uses the elementary inequality that $2ab \leq \eta a^2 + b^2/\eta$, with $\eta = \frac{1-\lambda^2}{2\lambda^2}$ for any $a, b \in \mathbb{R}$, and (v) holds since each f_i is L-smooth. Next, towards the last term in (5.41), we take the expectation

of (5.42) to obtain: $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{v}_t^i - \mathbf{v}_{t-1}^i\right\|^2\right] &\leq 3\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_t^i, \boldsymbol{\xi}_t^i) - \mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i)\right\|^2\right] + 3\beta^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] \\ &\quad + 3\beta^2\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_t^i) - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] \\ &\leq 3L^2\mathbb{E}\left[\left\|\mathbf{x}_t^i - \mathbf{x}_{t-1}^i\right\|^2\right] + 3\beta^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2\right] + 3\beta^2\nu_i^2, \end{aligned} \tag{5.45}$$

where (5.45) is due to the mean-squared smoothness and the bounded variance of each $\mathbf{g}_i$. Summing up (5.45) over i from 1 to n gives an upper bound on the last term in (5.41): $\forall t \geq 1$,

$$\lambda^2\mathbb{E}\left[\left\|\mathbf{v}_t - \mathbf{v}_{t-1}\right\|^2\right] \leq 3\lambda^2L^2\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}_{t-1}\right\|^2\right] + 3\lambda^2\beta^2\mathbb{E}\left[\left\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\right\|^2\right] + 3\lambda^2n\beta^2\overline{\nu}^2. \tag{5.46}$$

We now use (5.44) and (5.46) in (5.41) to obtain: $\forall t \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\right\|^2\right] &\leq \frac{1+\lambda^2}{2}\mathbb{E}\left[\left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2\right] + \frac{7\lambda^2L^2}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}_{t-1}\right\|^2\right] \\ &\quad + \frac{7\lambda^2\beta^2}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\right\|^2\right] + 3\lambda^2n\beta^2\overline{\nu}^2. \end{aligned} \tag{5.47}$$

Towards the second term in (5.47), we use (5.10) to obtain: $\forall t \geq 1$,

$$\begin{aligned}\left\|\mathbf{x}_t - \mathbf{x}_{t-1}\right\|^2 &= \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t + \mathbf{J}\mathbf{x}_t - \mathbf{J}\mathbf{x}_{t-1} + \mathbf{J}\mathbf{x}_{t-1} - \mathbf{x}_{t-1}\right\|^2 \\ &\overset{(i)}{\leq} 3\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + 3n\alpha^2\left\|\overline{\mathbf{v}}_{t-1}\right\|^2 + 3\left\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\right\|^2 \\ &\leq 6\lambda^2\alpha^2\left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2 + 3n\alpha^2\left\|\overline{\mathbf{v}}_{t-1}\right\|^2 + 9\left\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\right\|^2, \end{aligned} \tag{5.48}$$

where (i) uses the $\overline{\mathbf{x}}_t$-update in (5.5). Finally, we use (5.48) in (5.47) to obtain: $\forall t \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\right\|^2\right] \leq &\left(\frac{1+\lambda^2}{2} + \frac{42\lambda^4L^2\alpha^2}{1-\lambda^2}\right)\mathbb{E}\left[\left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2\right] + \frac{21\lambda^2nL^2\alpha^2}{1-\lambda^2}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_{t-1}\right\|^2\right] \\ &+ \frac{63\lambda^2L^2}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\right\|^2\right] + \frac{7\lambda^2\beta^2}{1-\lambda^2}\mathbb{E}\left[\left\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\right\|^2\right] + 3\lambda^2n\beta^2\overline{\nu}^2. \end{aligned}$$

The proof is completed by the fact that $\frac{1+\lambda^2}{2} + \frac{42\lambda^4L^2\alpha^2}{1-\lambda^2} \leq \frac{3+\lambda^2}{4}$ if $0 < \alpha \leq \frac{1-\lambda^2}{2\sqrt{42}\lambda^2L}$.

5.6.4 Proof of Lemma 5.5.6

5.6.4.1 Proof of Eq. (5.11)

We recursively apply the inequality on V_t from t to 0 to obtain: $\forall t \geq 1$,

$$\begin{aligned}V_t &\leq qV_{t-1} + qR_{t-1} + Q_t + C \\ &\leq q^2V_{t-2} + (q^2R_{t-2} + qR_{t-1}) + (qQ_{t-1} + Q_t) + (qC + C) \\ &\cdots \\ &\leq q^tV_0 + \sum_{i=0}^{t-1} q^{t-i}R_i + \sum_{i=1}^{t} q^{t-i}Q_i + C\sum_{i=0}^{t-1} q^i. \end{aligned} \tag{5.49}$$

Summing up (5.49) over t from 1 to T gives: $\forall T \geq 1$,

$$\begin{aligned}\sum_{t=0}^{T} V_t &\leq V_0 \sum_{t=0}^{T} q^t + \sum_{t=1}^{T}\sum_{i=0}^{t-1} q^{t-i} R_i + \sum_{t=1}^{T}\sum_{i=1}^{t} q^{t-i} Q_i + C\sum_{t=1}^{T}\sum_{i=0}^{t-1} q^i \\ &\leq V_0 \sum_{t=0}^{\infty} q^t + \sum_{t=0}^{T-1}\left(\sum_{i=0}^{\infty} q^i\right) R_t + \sum_{t=1}^{T}\left(\sum_{i=0}^{\infty} q^i\right) Q_t + C\sum_{t=1}^{T}\sum_{i=0}^{\infty} q^i,\end{aligned}$$

and the proof follows by $\sum_{i=0}^{\infty} q^i = (1-q)^{-1}$.

5.6.4.2 Proof of Eq. (5.12)

We recursively apply the inequality on V_t from $t+1$ to 1 to obtain: $\forall t \geq 1$,

$$\begin{aligned}V_{t+1} &\leq qV_t + R_{t-1} + C \\ &\leq q^2 V_{t-1} + (qR_{t-2} + R_{t-1}) + (qC + C) \\ &\cdots \\ &\leq q^t V_1 + \sum_{i=0}^{t-1} q^{t-1-i} R_i + C\sum_{i=0}^{t-1} q^i. \end{aligned} \tag{5.50}$$

We sum up (5.50) over t from 1 to $T-1$ to obtain: $\forall T \geq 2$,

$$\begin{aligned}\sum_{t=0}^{T-1} V_{t+1} &\leq V_1 \sum_{t=0}^{T-1} q^t + \sum_{t=1}^{T-1}\sum_{i=0}^{t-1} q^{t-1-i} R_i + C\sum_{t=1}^{T-1}\sum_{i=0}^{t-1} q^i \\ &\leq V_1 \sum_{t=0}^{\infty} q^t + \sum_{t=0}^{T-2}\left(\sum_{i=0}^{\infty} q^i\right) R_t + C\sum_{t=1}^{T-1}\sum_{i=0}^{\infty} q^i,\end{aligned}$$

and the proof follows by $\sum_{i=0}^{\infty} q^i = (1-q)^{-1}$.

5.6.5 Proof of Lemma 5.5.7

5.6.5.1 Proof of Eq. (5.13)

We first observe that $\frac{1}{1-(1-\beta)^2} \leq \frac{1}{\beta}$ for $\beta \in (0,1)$. Applying (5.11) to (5.7) gives: $\forall T \geq 1$,

$$\begin{aligned}&\sum_{t=0}^{T} \mathbb{E}\left[\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2\right] \\ &\leq \frac{\mathbb{E}\left[\|\overline{\mathbf{v}}_0 - \overline{\nabla \mathbf{f}}(\mathbf{x}_0)\|^2\right]}{\beta} + \frac{6L^2\alpha^2}{n\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{6L^2}{n^2\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{J}\mathbf{x}_{t+1}\right\|^2 + \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + \frac{2\beta\overline{\nu}^2 T}{n} \\ &\leq \frac{\mathbb{E}\left[\|\overline{\mathbf{v}}_0 - \overline{\nabla \mathbf{f}}(\mathbf{x}_0)\|^2\right]}{\beta} + \frac{6L^2\alpha^2}{n\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{12L^2}{n^2\beta}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + \frac{2\beta\overline{\nu}^2 T}{n}.\end{aligned} \tag{5.51}$$

Towards the first term in (5.51), we observe that

$$\mathbb{E}\left[\left\|\overline{\mathbf{v}}_0 - \overline{\nabla \mathbf{f}}(\mathbf{x}_0)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b_0}\sum_{r=1}^{b_0}\left(\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right)\right\|^2\right]$$
$$\overset{(i)}{=} \frac{1}{n^2 b_0^2}\sum_{i=1}^{n}\sum_{r=1}^{b_0}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right\|^2\right] \leq \frac{\overline{\nu}^2}{n b_0}, \tag{5.52}$$

where (i) follows from a similar line of arguments in (5.33). Then (5.13) follows from using (5.52) in (5.51).

5.6.5.2 Proof of Eq. (5.14)

We apply (5.11) to (5.8) to obtain: $\forall T \geq 1$,

$$\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{x}_t)\right\|^2\right]$$
$$\leq \frac{\mathbb{E}\left[\left\|\mathbf{v}_0 - \nabla \mathbf{f}(\mathbf{x}_0)\right\|^2\right]}{\beta} + \frac{6nL^2\alpha^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{6L^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{J}\mathbf{x}_{t+1}\right\|^2 + \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + 2n\beta\overline{\nu}^2 T$$
$$\leq \frac{\mathbb{E}\left[\left\|\mathbf{v}_0 - \nabla \mathbf{f}(\mathbf{x}_0)\right\|^2\right]}{\beta} + \frac{6nL^2\alpha^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t\right\|^2\right] + \frac{12L^2}{\beta}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + 2n\beta\overline{\nu}^2 T. \tag{5.53}$$

In (5.53), we observe that

$$\mathbb{E}\left[\left\|\mathbf{v}_0 - \nabla \mathbf{f}(\mathbf{x}_0)\right\|^2\right] = \sum_{i=1}^{n}\mathbb{E}\left[\left\|\frac{1}{b_0}\sum_{r=1}^{b_0}\left(\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right)\right\|^2\right]$$
$$\overset{(i)}{=} \frac{1}{b_0^2}\sum_{i=1}^{n}\sum_{r=1}^{b_0}\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_0^i, \boldsymbol{\xi}_{0,r}^i) - \nabla f_i(\mathbf{x}_0^i)\right\|^2\right] \leq \frac{n\overline{\nu}^2}{b_0}, \tag{5.54}$$

where (i) follows from a similar line of arguments in (5.33). Then (5.14) follows from using (5.54) in (5.53).

5.6.6 Proof of Lemma 5.5.8

We recall that $\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\| = 0$, since it is assumed without generality that $\mathbf{x}_0^i = \mathbf{x}_0^j$ for any $i, j \in \mathcal{V}$. Applying (5.11) to (5.9) yields: $\forall T \geq 1$,

$$\sum_{t=0}^{T}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 \leq \frac{4\lambda^2\alpha^2}{(1-\lambda^2)^2}\sum_{t=1}^{T}\left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2. \tag{5.55}$$

To further bound $\sum_{t=1}^{T}\|\mathbf{y}_t-\mathbf{J}\mathbf{y}_t\|^2$, we apply (5.12) in Lemma 5.5.5(b) to obtain: if $0<\alpha\leq\frac{1-\lambda^2}{2\sqrt{42\lambda^2}L}$, then $\forall T\geq 2$,

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{y}_t-\mathbf{J}\mathbf{y}_t\|^2\right]\\
&\leq\frac{4\mathbb{E}\left[\|\mathbf{y}_1-\mathbf{J}\mathbf{y}_1\|^2\right]}{1-\lambda^2}+\frac{84\lambda^2nL^2\alpha^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right]+\frac{252\lambda^2L^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{J}\mathbf{x}_t\|^2\right]\\
&\quad+\frac{28\lambda^2\beta^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{x}_t)\|^2\right]+\frac{12\lambda^2n\beta^2\overline{\nu}^2T}{1-\lambda^2}\\
&\leq\frac{84\lambda^2nL^2\alpha^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right]+\frac{252\lambda^2L^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{J}\mathbf{x}_t\|^2\right]\\
&\quad+\frac{28\lambda^2\beta^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{x}_t)\|^2\right]+\frac{12\lambda^2n\beta^2\overline{\nu}^2T}{1-\lambda^2}+\frac{4\lambda^2\left\|\nabla\mathbf{f}\left(\mathbf{x}_0\right)\right\|^2}{1-\lambda^2}+\frac{4\lambda^2n\overline{\nu}^2}{(1-\lambda^2)b_0},
\end{aligned}\tag{5.56}$$

where the last inequality is due to Lemma 5.5.5(a). To proceed, we use (5.14), an upper bound on $\sum_t\mathbb{E}\left[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{x}_t)\|^2\right]$, in (5.56) to obtain: if $0<\alpha\leq\frac{1-\lambda^2}{2\sqrt{42\lambda^2}L}$ and $\beta\in(0,1)$, then $\forall T\geq 2$,

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{y}_t-\mathbf{J}\mathbf{y}_t\|^2\right]&\leq\frac{252\lambda^2nL^2\alpha^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right]+\frac{588\lambda^2L^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{J}\mathbf{x}_t\|^2\right]\\
&\quad+\frac{28\lambda^2n\beta\overline{\nu}^2}{(1-\lambda^2)^2b_0}+\frac{56\lambda^2n\beta^3\overline{\nu}^2T}{(1-\lambda^2)^2}+\frac{12\lambda^2n\beta^2\overline{\nu}^2T}{1-\lambda^2}+\frac{4\lambda^2\left\|\nabla\mathbf{f}\left(\mathbf{x}_0\right)\right\|^2}{1-\lambda^2}+\frac{4\lambda^2n\overline{\nu}^2}{(1-\lambda^2)b_0}\\
&=\frac{252\lambda^2nL^2\alpha^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right]+\frac{588\lambda^2L^2}{(1-\lambda^2)^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{J}\mathbf{x}_t\|^2\right]\\
&\quad+\left(\frac{7\beta}{1-\lambda^2}+1\right)\frac{4\lambda^2n\overline{\nu}^2}{(1-\lambda^2)b_0}+\left(\frac{14\beta}{1-\lambda^2}+3\right)\frac{4\lambda^2n\beta^2\overline{\nu}^2T}{1-\lambda^2}+\frac{4\lambda^2\left\|\nabla\mathbf{f}\left(\mathbf{x}_0\right)\right\|^2}{1-\lambda^2}.
\end{aligned}\tag{5.57}$$

Finally, we use (5.57) in (5.55) to obtain: $\forall T\geq 2$,

$$\begin{aligned}
\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{J}\mathbf{x}_t\|^2\right]&\leq\frac{1008\lambda^4nL^2\alpha^4}{(1-\lambda^2)^4}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right]+\frac{2352\lambda^4L^2\alpha^2}{(1-\lambda^2)^4}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{J}\mathbf{x}_t\|^2\right]\\
&\quad+\left(\frac{7\beta}{1-\lambda^2}+1\right)\frac{16\lambda^4n\overline{\nu}^2\alpha^2}{(1-\lambda^2)^3b_0}+\left(\frac{14\beta}{1-\lambda^2}+3\right)\frac{16\lambda^4n\beta^2\overline{\nu}^2\alpha^2T}{(1-\lambda^2)^3}+\frac{16\lambda^4\left\|\nabla\mathbf{f}\left(\mathbf{x}_0\right)\right\|^2\alpha^2}{(1-\lambda^2)^3},
\end{aligned}$$

which may be written equivalently as

$$\begin{aligned}
\left(1-\frac{2352\lambda^4L^2\alpha^2}{(1-\lambda^2)^4}\right)\sum_{t=0}^{T}\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{J}\mathbf{x}_t\|^2\right]&\leq\frac{1008\lambda^4nL^2\alpha^4}{(1-\lambda^2)^4}\sum_{t=0}^{T-2}\mathbb{E}\left[\|\overline{\mathbf{v}}_t\|^2\right]+\left(\frac{7\beta}{1-\lambda^2}+1\right)\frac{16\lambda^4n\overline{\nu}^2\alpha^2}{(1-\lambda^2)^3b_0}\\
&\quad+\left(\frac{14\beta}{1-\lambda^2}+3\right)\frac{16\lambda^4n\beta^2\overline{\nu}^2\alpha^2T}{(1-\lambda^2)^3}+\frac{16\lambda^4\left\|\nabla\mathbf{f}\left(\mathbf{x}_0\right)\right\|^2\alpha^2}{(1-\lambda^2)^3}.
\end{aligned}\tag{5.58}$$

We observe in (5.58) that $\frac{2352\lambda^4L^2\alpha^2}{(1-\lambda^2)^4}\leq\frac{1}{2}$ if $0<\alpha\leq\frac{(1-\lambda^2)^2}{70\lambda^2L}$, and the proof follows.

5.7 Conclusion

In this chapter, we propose `GT-HSGD`, a stochastic variance-reduced gradient algorithm as an instance of the `GT-VR` framework developed in Chapter 2, for decentralized non-convex expected risk minimization problems with mean-squared smoothness. It is shown that `GT-HSGD` achieves an improved oracle complexity compared to the existing decentralized stochastic gradient methods. Furthermore, we show that the oracle complexity of `GT-HSGD`, when the required error tolerance is small enough, reduces to $O(\epsilon^{-3})$, which is independent of the network topology and matches that of the centralized optimal methods for this problem class. To the best of our knowledge, this is the first such result in the literature.

Chapter 6

Decentralized Stochastic Non-Convex Composite Optimization

In this chapter, we consider decentralized non-convex composite problems, where the goal of the networked nodes is to find a first-order stationary point of the average of local, smooth, possibly non-convex risk functions plus an extended valued, convex, possibly non-differentiable regularizer. This non-convex non-smooth composite problem may be viewed as a generalization of the problems considered in the previus chapters. To tackle this general formulation, we develop a unified stochastic gradient tracking framework, `ProxGT`, that allows flexible constructions of local stochastic (variance-reduced) gradient estimators. For definiteness, we construct instantiations of `ProxGT` by specifying appropriate local estimators for several problem classes of interest. For each problem class, an instance of `ProxGT` achieves gradient and communication complexities that match that of the corresponding centralized optimal methods. Several intermediate technical results in the convergence analysis are of independent interest. Numerical simulation results are included to demonstrate our theoretical claims.[1]

6.1 Introduction

Decentralized optimization, also known as distributed optimization over graphs, is a general parallel computation model for minimizing a sum of cost functions distributed over a network of nodes without a central coordinator [33]. This cooperative minimization paradigm, built upon local communication and computation, has numerous applications in estimation and learning problems that arise in multi-agent systems [21,31,37]. In particular, the sparse and localized peer-to-peer information exchange pattern in decentralized networks substantially reduces the communication overhead on the parameter server in the centralized networks, thus

[1] The content presented in this chapter can be partially found in [164].

making decentralized optimization algorithms especially appealing in large-scale data analytics and machine learning tasks [2, 3, 29, 40, 41, 165].

In this chapter, we consider the following decentralized *non-convex non-smooth composite* optimization problem defined over a network of n nodes:

$$\min_{\mathbf{x}\in\mathbb{R}^p} \Psi(\mathbf{x}) := F(\mathbf{x}) + h(\mathbf{x}), \quad F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}). \tag{6.1}$$

Here, each $f_i : \mathbb{R}^p \to \mathbb{R}$ is smooth, possibly non-convex, and is only locally accessible by node i, while $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is convex, possibly non-differentiable, and is commonly known by all nodes. In particular, each f_i is a cost function associated with local data at node i, while h serves as a regularization term that is often used to impose additional problem structure such as convex constraints and/or sparsity; common examples of h include the indicator function of a convex set or the ℓ_1-norm. The communication over the networked nodes is abstracted as a directed graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} := \{1, \cdots, n\}$ denotes the set of node indices and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ collects ordered pairs (i, r), $i, r \in \mathcal{V}$, such that node r sends information to node i. Our focus in this chapter is on the following formulations of the local costs $\{f_i\}_{i=1}^n$ that frequently appear in the context of machine learning [7]:

- **Expected risk.** In this case, each f_i in Problem (6.1) is defined as

$$f_i(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\xi}_i\sim\mathcal{D}_i}[G_i(\mathbf{x}, \boldsymbol{\xi}_i)], \tag{6.2}$$

  where $\boldsymbol{\xi}_i$ is a random data vector supported on $\Xi_i \subseteq \mathbb{R}^q$ with some unknown probability distribution $\mathcal{D}_i$ and $G_i : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is a Borel function. The stochastic formulation (6.2) often corresponds to *online* scenarios such that samples are generated from the underlying data stream in real time at each node i, in order to construct stochastic approximation of f_i for the subsequent optimization procedure [166].

- **Empirical risk.** We are also concerned with a special case of (6.2), i.e., when $\boldsymbol{\xi}_i$ has a finite support set $\Xi_i := \{\boldsymbol{\xi}_{i,(1)}, \cdots, \boldsymbol{\xi}_{i,(m)}\}$ for some $m \geq 1$ and each f_i takes the deterministic form of

$$f_i(\mathbf{x}) := \frac{1}{m}\sum_{s=1}^{m} G_i\big(\mathbf{x}, \boldsymbol{\xi}_{i,(s)}\big). \tag{6.3}$$

  As an alternate viewpoint, the formulation (6.3) may be considered as the sample average approximation of (6.2), where $\{\boldsymbol{\xi}_{i,(1)}, \cdots, \boldsymbol{\xi}_{i,(m)}\}$ take the role of *offline* samples generated from the distribution $\mathcal{D}_i$ [167]. We are particularly interested in modern-day big-data scenarios, where the local sample size m is very large and thus stochastic gradient methods are often preferable over exact gradient ones that use the entire local data per update.

The above formulations are quite general and have found applications in, e.g., sparse non-convex linear models [146], principle component analysis [129], and matrix factorization [168]. Our goal in this chapter is thus on the design and analysis of efficient decentralized *stochastic* gradient algorithms to find an ϵ-stationary point of the global *non-convex non-smooth composite* function Ψ in Problem (6.1) under expected risk (6.2) or empirical risk (6.3).

6.1.1 Related work

The last decade has witnessed a growing literature in the area of decentralized optimization; see, e.g., survey articles [14, 27]. For convex composite problems, we refer the readers to, e.g., [74, 75, 169–172] and the references therein. On the other hand, the work on decentralized methods for non-convex non-smooth composite problems is fairly limited. In the following, we review the existing results that are closely related to Problem (6.1) under either (6.2) or (6.3).

The first algorithmic framework for decentralized non-convex composite problems was proposed by [55], where h is handled in a successive convex approximation scheme with the help of gradient tracking. Reference [173] presents decentralized proximal gradient descent which tackles h via proximal mapping. These works [55, 173], however, require the gradient of F and the subdifferential of h to be uniformly bounded. These boundedness assumptions are removed in [129, 174], where unbalanced directed graphs and compression are also considered respectively. A decentralized Frank-Wolfe method is proposed in [72] to handle the case where h is the indicator function of a convex compact set. We note that the aforementioned results [55, 72, 129, 173, 174] are exact gradient methods, which are in general not applicable to the expected risk (6.2) and also may not be sample-efficient in the empirical risk setting (6.3) when the local data size m is relatively large. Towards stochastic gradient methods, [175] analyzes a projected `DSGD` method for problems with compact inequality constraints. Reference [42] establishes the asymptotic convergence of `DSGD` for a family of non-convex non-smooth coercive functions. A recent work [176] presents `SPPDM`, a decentralized stochastic proximal primal-dual method, and provides related convergence guarantees under the assumption that the epigraph of h is a *polyhedral* set.

To the best of our knowledge, the decentralized stochastic optimization literature *lacks non-asymptotic gradient and communication complexity results for the non-convex non-smooth composite problem under a general convex, possibly non-differentiable regularizer h.* We address this gap in this chapter.

Table 6.1: A summary of the gradient and communication complexities of the instances of `ProxGT` studied in this chapter for finding an ϵ-stationary point of the global composite function Ψ over an undirected network. In the table, n is the number of the nodes, $(1-\lambda_*) \in (0,1]$ is the spectral gap of the weight matrix associated with the network, L is the smoothness parameter for the risk functions, Δ is the function value gap, ν^2 is the stochastic gradient variance under the expected risk, m is the local sample size under the empirical risk. The MSS column specifies whether the convergence of the algorithm in question requires the mean-squared smoothness assumption.

Algorithm	**Sample Complexity at Each Node**	**Communication Complexity**	**MSS**	**Remarks**
`ProxGT-SA`	$\mathcal{O}\left(\frac{L\Delta\nu^2}{n\epsilon^4}\right)$	$\mathcal{O}\left(\frac{L\Delta}{\epsilon^2}\cdot\frac{\log n}{\sqrt{1-\lambda_*}}\right)$	✗	Population Risk (6.2)
`ProxGT-SR-O`	$\mathcal{O}\left(\frac{L\Delta\nu}{n\epsilon^3}+\frac{\nu^2}{n\epsilon^2}\right)$	$\mathcal{O}\left(\left(\frac{L\Delta}{\epsilon^2}+\frac{\nu}{\epsilon}\right)\cdot\frac{\log n}{\sqrt{1-\lambda_*}}\right)$	✓	Population Risk (6.2)
`ProxGT-SR-E`	$\mathcal{O}\left(\frac{L\Delta}{\epsilon^2}\max\left\{\sqrt{\frac{m}{n}},1\right\}+\max\left\{m,\sqrt{nm}\right\}\right)$	$\mathcal{O}\left(\left(\frac{L\Delta}{\epsilon^2}+\sqrt{nm}\right)\cdot\frac{1}{\sqrt{1-\lambda_*}}\right)$	✓	Empirical Risk (6.3)

6.1.2 Main contributions

We develop `ProxGT`, a unified stochastic proximal gradient tracking framework for designing and analyzing decentralized methods for the general non-convex non-smooth composite problem. `ProxGT` allows flexible construction of local gradient estimators, where a suitable one may be chosen in light of the underlying problem specifications and practical applications. We highlight our main contributions in the following.

- **Algorithms**. We present three instances of the proposed `ProxGT` framework. For the general expected risk, we develop `ProxGT-SA` by using the minibatch stochastic approximation technique [177]. Leveraging `SARAH/SPIDER` type recursive variance reduction schemes [48–50], we provide two accelerated algorithms, named `ProxGT-SR-O` and `ProxGT-SR-E`, for the population and empirical risk respectively that outperform `ProxGT-SA` under a mean-squared smoothness property [149].

- **Gradient and communication complexity results**. We establish non-asymptotic gradient and communication complexities of the proposed `ProxGT-SA`, `ProxGT-SR-O`, and `ProxGT-SR-E` algorithms to find an ϵ-stationary solution of the non-convex non-smooth composite problem; see Table 6.1 for a summary. Remarkably, these sample complexities at each node are network topology-independent and are n times smaller than that of the centralized *optimal* algorithms [48, 50, 177] implemented on a single node for the corresponding problem classes. In other words, the proposed methods achieve a topology-independent linear speedup compared to their respective optimal centralized counterparts.

- **Special cases**. For the special case $h = 0$, it is worth emphasizing that `ProxGT-SR-E` and `ProxGT-SR-O` also constitute improvements over the state-of-the-art decentralized variance-reduced methods `GT-SARAH` [141] and `GT-HSGD` [178] for smooth problems in the following sense. For the empirical risk, `GT-SARAH` attains

the optimal centralized gradient complexity when the local sample size m is large enough. Similarly, for the expected risk, the gradient complexity of `GT-HSGD` is optimal when the required accuracy is small enough. `ProxGT-SR-O` and `ProxGT-SR-E` relax these regime restrictions and achieve improved communication complexities simultaneously, by performing multiple (accelerated) consensus updates per iteration with proper mini-batches.

- **Analysis techniques**. We establish a new *stochastic gradient mapping descent inequality* and a new *consensus error bound* for the non-convex non-smooth composite problem. Their proofs are novel and substantially different from their counterparts in decentralized smooth optimization, e.g., [135, 141, 178, 179], due to the nonlinear coupling of the proximal mapping, gradient noise, consensus errors, and non-convexity of the risks. We emphasize that these intermediate technical results are of independent interest and are instrumental in analyzing other methods based on similar principles, such as proximal `DSGD` and its variants. This is because these results hold true regardless of the underlying gradient estimation procedure. Finally, we note that the convergence analyses of `ProxGT-SA`, `ProxGT-SR-O`, and `ProxGT-SR-E` are developed in a unified manner and can be used to analyze other instances of the `ProxGT` framework.

The set of positive real numbers is denoted by $\mathbb{R}^+$. For an integer $z \geq 1$, we denote $[z] := \{1, \cdots, z\}$. We use lowercase bold letters to denote vectors and uppercase bold letters to denote matrices. The $d \times d$ identity matrix is denoted by $\mathbf{I}_d$, while the d-dimensional column vectors of all ones and zeros are represented by $\mathbf{1}_d$ and $\mathbf{0}_d$, respectively. For a matrix $\mathbf{X} \in \mathbb{R}^{d\times d}$, its (i,r)-th entry is denoted by $[\mathbf{X}]_{i,r}$. The Kronecker product of two matrices is denoted by $\otimes$. The ℓ_2-norm of a vector or the spectral norm of a matrix is denoted by $\|\cdot\|$, while the ℓ_1-norm of a vector is denoted by $\|\cdot\|_1$. For an extended valued function $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, we denote $\mathrm{dom}(h) := \{\mathbf{x} : h(\mathbf{x}) < +\infty\}$, and h is said to be proper if $\mathrm{dom}(h)$ is nonempty. For $\mathbf{x} \in \mathrm{dom}(h)$, we denote the subdifferential of h at $\mathbf{x}$ by $\partial h(\mathbf{x})$. The proximal mapping of h is defined as

$$\mathrm{prox}_h(\mathbf{x}) := \mathrm{argmin}_{\mathbf{u}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 + h(\mathbf{u}) \right\}. \tag{6.4}$$

Given a σ-algebra $\mathcal{H}$ and a random vector $\mathbf{x}$, we write $\mathbf{x} \in \mathcal{H}$ if $\mathbf{x}$ is $\mathcal{H}$-measurable. We use $\sigma(\cdot)$ to denote the generated σ-algebra.

The remainder of this chapter is organized as follows. Section 6.2 formulates the problems. Section 6.3 develops the proposed algorithmic framework and its instances of interest in this paper. Section 6.4 presents the main convergence results of the proposed algorithms and discuss their implications. Numerical illustrations are presented in Section 6.5. The convergence analysis outline of the proposed algorithms is provided in Section 6.6, while the detailed proofs and derivations are presented in Section 6.7.

6.2 Problem formulation

In this section, we formulate the optimization and network models of interest in this chapter.

6.2.1 The non-convex non-smooth composite model

Throughout this chapter, we make the following assumption on the objective functions.

Assumption 6.2.1 (**Functions**). *In Problem (6.1), the following statements hold:*

(a) $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ *is proper, closed, and convex;*

(b) Each $f_i : \mathbb{R}^p \to \mathbb{R}$ *is* L*-smooth, i.e.,* $\exists L \in \mathbb{R}^+$*, s.t.* $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$*,* $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$*;*

(c) Ψ *is bounded below, i.e.,* $\underline{\Psi} := \inf_{\mathbf{x}\mathbb{R}^p} \Psi(\mathbf{x}) > -\infty$.

Assumption 6.2.1 characterizes the standard *non-convex non-smooth composite model* [10], where a point $\widehat{\mathbf{x}} \in \mathrm{dom}(h)$ is said to be *stationary* for Problem (6.1) if

$$-\nabla F(\widehat{\mathbf{x}}) \in \partial h(\widehat{\mathbf{x}}). \tag{6.5}$$

A simple example of an extended valued function h that satisfies Assumption 6.2.1(a) is the indicator of a nonempty, closed, and convex set in $\mathbb{R}^p$.

Remark 6.2.1. It is shown in [10] that the stationary condition (6.5) is a necessary condition for a point $\widehat{\mathbf{x}}$ to be a local optimal solution of Problem (6.1).

With the help of the proximal mapping (6.4), it can be shown that the stationarity condition (6.5) is equivalent to a fixed point equation [10], i.e., $\widehat{\mathbf{x}} \in \mathbb{R}^p$ is stationary for Problem (6.1) if and only if

$$\widehat{\mathbf{x}} = \mathrm{prox}_{\alpha h}\big(\widehat{\mathbf{x}} - \alpha \nabla F(\widehat{\mathbf{x}})\big), \qquad \forall \alpha > 0. \tag{6.6}$$

In view of (6.6), we define the *gradient mapping* [8, 10] for Problem (6.1): $\forall \mathbf{x} \in \mathrm{dom}(h)$,

$$\mathbf{s}(\mathbf{x}) := \frac{1}{\alpha}\Big(\mathbf{x} - \mathrm{prox}_{\alpha h}\big(\mathbf{x} - \alpha \nabla F(\mathbf{x})\big)\Big), \tag{6.7}$$

where $\alpha > 0$. We note that the gradient mapping $\mathbf{s}(\mathbf{x})$ can be viewed as a generalized gradient of Ψ at $\mathbf{x}$ in the sense that $\mathbf{s}(\mathbf{x}) = \nabla F(\mathbf{x})$ if $h = 0$. The size of $\mathbf{s}(\cdot)$ thus serves as a natural measure for the approximate stationarity of a solution [8, 10].

Definition 6.2.1 (ϵ**-stationarity**). *Under Assumption 6.2.1, we say a random vector* $\mathbf{x} \in$ *dom*(h) *is an* ϵ*-stationary solution for Problem (6.1) if* $\mathbb{E}[\|\mathbf{s}(\mathbf{x})\|] \leq \epsilon$*, where* $\sigma(\cdot)$ *is defined in (6.7).*

6.2.2 The network model

We consider the following assumption on the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which characterizes the decentralized communication between the networked nodes.

Assumption 6.2.2 (**Network**). *The directed network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is strongly connected. Moreover, there exists a weight matrix $\mathbf{W}_* \in \mathbb{R}^{n\times n}$ associated with $\mathcal{G}$ that satisfies the following conditions:*

(a) $[\mathbf{W}_*]_{i,r} > 0$, *if* $(i,r) \in \mathcal{E}$;

(b) $[\mathbf{W}_*]_{i,r} = 0$, *if* $(i,r) \notin \mathcal{E}$;

(c) $\mathbf{W}_*\mathbf{1}_n = \mathbf{W}_*^\top \mathbf{1}_n = \mathbf{1}_n$.

Under Assumption 6.2.2, it is well known that the consensus weight matrix $\mathbf{W}_*$ is primitive and doubly stochastic [27, 36], i.e.,

$$\lambda_* := \big\|\mathbf{W}_* - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\big\| \in [0,1). \tag{6.8}$$

We refer $(1-\lambda_*) \in (0,1]$ as the *spectral gap* of $\mathcal{G}$ which characterizes the connectivity of the network.

6.2.3 Stochastic gradient models

We make a blanket assumption that each node i at every iteration t is able to obtain i.i.d. minibatch samples $\{\boldsymbol{\xi}_{i,s}^t : s \in [b_t]\}$ for the local random data vector $\boldsymbol{\xi}_i$. The induced natural filtration is given by, $\forall t \geq 2$,

$$\mathcal{F}_t := \sigma\big(\boldsymbol{\xi}_{i,s}^r : \forall i \in \mathcal{V}, s \in [b_r], 1 \leq r \leq t-1\big), \tag{6.9}$$

while $\mathcal{F}_1$ is the trivial σ-algebra. Intuitively, the filtration $\mathcal{F}_t$ collects the historical information of an algorithm that constantly samples $\boldsymbol{\xi}_i$ up to iteration t. We require that the stochastic gradient $\nabla G(\cdot, \boldsymbol{\xi}_i)$ is conditionally unbiased.

Assumption 6.2.3 (**Unbiasedness**). $\forall i \in \mathcal{V}$, $\forall t \geq 1$, $\forall \mathbf{x} \in \mathcal{F}_t$, *we have* $\mathbb{E}\big[\nabla G_i(\mathbf{x}, \boldsymbol{\xi}_i)|\mathcal{F}_t\big] = \nabla f_i(\mathbf{x})$.

Remark 6.2.2. Under the empirical risk formulation (6.3), Assumption 6.2.3 amounts to uniform sampling indices at random from $[m]$ at each node.

We next consider a bounded variance assumption [8] for the stochastic gradient $\nabla G(\cdot, \boldsymbol{\xi}_i)$, which will be used in the expected risk setting (6.2).

Assumption 6.2.4 (**Bounded Variance**). *Let $\nu_i \in \mathbb{R}^+$, $\forall i \in \mathcal{V}$. We have* $\mathbb{E}\big[\|\nabla G_i(\mathbf{x}, \boldsymbol{\xi}_i) - \nabla f_i(\mathbf{x})\|^2|\mathcal{F}_t\big] \leq \nu_i^2$, $\forall t \geq 1$, $\forall \mathbf{x} \in \mathcal{F}_t$, $\forall i \in \mathcal{V}$; $\nu^2 := \frac{1}{n}\sum_{i=1}^n \nu_i^2$.

We are also interested in the case when the stochastic gradients further satisfy a mean-squared smoothness property [149], which is often satisfied by many machine learning models [8].

Assumption 6.2.5 (**Mean-Squared Smoothness**). *Let $L \in \mathbb{R}^+$. For the expected risk (6.2), we have*

$$\mathbb{E}\big[\|\nabla G_i(\mathbf{x}, \boldsymbol{\xi}_i) - \nabla G_i(\mathbf{y}, \boldsymbol{\xi}_i)\|^2\big] \leq L^2 \mathbb{E}\big[\|\mathbf{x} - \mathbf{y}\|^2\big],$$

for all $i \in \mathcal{V}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. In the case of empirical risk (6.3), the above statement reduces to

$$\frac{1}{m} \sum_{s=1}^{m} \big\|\nabla G_i(\mathbf{x}, \boldsymbol{\xi}_{i,(s)}) - \nabla G_i(\mathbf{y}, \boldsymbol{\xi}_{i,(s)})\big\|^2 \leq L^2 \|\mathbf{x} - \mathbf{y}\|^2,$$

for all $i \in \mathcal{V}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$.

6.3 Algorithm development

A popular centralized fixed-point method to solve (6.6) is the proximal gradient descent method [10]:

$$\mathbf{x}_{t+1} = \mathrm{prox}_{\alpha h}\big(\mathbf{x}_t - \alpha \nabla F(\mathbf{x}_t)\big), \qquad \forall t \geq 1, \tag{6.10}$$

where $\alpha > 0$. However, (6.10) cannot be directly implemented in a decentralized manner. The main challenge lies in the fact that the global gradient ∇F cannot be computed via one-shot aggregation of local gradient information in decentralized networks. Moreover, the local gradients $\{\nabla f_i\}_{i=1}^n$ are often significantly different due to the heterogeneous data across the nodes, making the classical gradient consensus approaches [27] less effective especially in the non-convex settings [141]. One popular technique to overcome these issues is gradient tracking [55, 65], which has been adopted in several decentralized stochastic gradient methods for smooth non-convex problems; see, e.g., [4, 135, 150, 178, 179]. Inspired by these works, we propose a general proximal stochastic gradient tracking framework, termed as `ProxGT`, to tackle the non-convex non-smooth composite Problem (6.1).

6.3.1 A generic algorithmic procedure

We now describe the proposed `ProxGT` framework. At every iteration t, each node i in the network retains three local variables $\mathbf{x}_t^i$, $\mathbf{v}_t^i$, and $\mathbf{y}_t^i$, all in $\mathbb{R}^p$, where $\mathbf{x}_t^i$ approximates a stationary point of Problem (6.1), $\mathbf{v}_t^i$ estimates the local exact gradient $\nabla f_i(\mathbf{x}_t^i)$ from the samples generated for $\boldsymbol{\xi}_i$, and $\mathbf{y}_t^i$ tracks the global gradient $\nabla F(\mathbf{x}_t^i)$ via a stochastic gradient tracking type update [67] from the local gradient estimates $\{\mathbf{v}_t^i\}_{i=1}^n$. With the global gradient tracker $\mathbf{y}_t^i$ at hand, each node i performs a local inexact fixed point update for (6.6):

$$\mathbf{z}_{t+1}^i := \mathrm{prox}_{\alpha h}\big(\mathbf{x}_t^i - \alpha \mathbf{y}_{t+1}^i\big),$$

Algorithm 7 `ProxGT` for Problem (6.1)

Require: $\mathbf{x}_1 = \mathbf{1}_n \otimes \overline{\mathbf{x}}_1$; K; α; $\mathbf{y}_1 = \mathbf{0}_{np}$; $\mathbf{v}_0 = \mathbf{0}_{np}$.

1: **for** $t = 1, \cdots, T$ **do**
2: Generate an estimator $\mathbf{v}_t^i$ of $\nabla f_i(\mathbf{x}_t^i)$, $\forall i$.
3: Tracking: $\mathbf{y}_{t+1} = \mathbf{W}^K\big(\mathbf{y}_t + \mathbf{v}_t - \mathbf{v}_{t-1}\big)$.
4: Prox-Descent: $\mathbf{z}_{t+1}^i = \mathrm{prox}_{\alpha h}\big(\mathbf{x}_t^i - \alpha \mathbf{y}_{t+1}^i\big)$, $\forall i$.
5: Consensus: $\mathbf{x}_{t+1} = \mathbf{W}^K \mathbf{z}_{t+1}$.
6: **end for**

where $\alpha > 0$ is the step-size. The local solution $\mathbf{x}_{t+1}^i$ at the next iteration is then updated by performing consensus on the intermediate variables $\{\mathbf{z}_{t+1}^i\}_{i=1}^n$ over the network. For the ease of presentation, we define

$$\mathbf{W} := \mathbf{W}_* \otimes \mathbf{I}_p.$$

and the global variables $\mathbf{x}_t$, $\mathbf{v}_t$, $\mathbf{y}_t$, $\mathbf{z}_t$ which concatenate their corresponding local variables, i.e.,

$$\mathbf{x}_t = \begin{bmatrix} \mathbf{x}_t^1 \\ \vdots \\ \mathbf{x}_t^n \end{bmatrix}, \quad \mathbf{v}_t = \begin{bmatrix} \mathbf{v}_t^1 \\ \vdots \\ \mathbf{v}_t^n \end{bmatrix}, \quad \mathbf{y}_t = \begin{bmatrix} \mathbf{y}_t^1 \\ \vdots \\ \mathbf{y}_t^n \end{bmatrix}, \quad \mathbf{z}_t = \begin{bmatrix} \mathbf{z}_t^1 \\ \vdots \\ \mathbf{z}_t^n \end{bmatrix},$$

all in $\mathbb{R}^{np}$. With the help of these notations, we formally present `ProxGT` in Algorithm 7 from a global view.

Remark 6.3.1. In Algorithm 7, the decentralized propagation and averaging of local variables over the network appear as matrix-vector products, while the node-wise implementation of `ProxGT` can be obtained accordingly.

Remark 6.3.2. We note that $\mathbf{W}^K$ leads to K decentralized averaging step(s) over K rounds of communication in the corresponding update. This multi-consensus update with an appropriately chosen K is often helpful to achieve faster convergence in the corresponding algorithms [120, 171, 180].

It is straightforward to show by induction that the $\mathbf{y}$-update in Algorithm 7 satisfies an important dynamic tracking property [51]:

$$\frac{1}{n}\sum_{i=1}^n \mathbf{y}_{t+1}^i = \frac{1}{n}\sum_{i=1}^n \mathbf{v}_t^i, \qquad \forall t \geq 1. \tag{6.11}$$

In view of (6.11) and the recursion of Algorithm 7, it is expected that each $\mathbf{y}_t^i$ approaches $\frac{1}{n}\sum_{i=1}^n \mathbf{v}_t^i$ and thus asymptotically tracks the global gradient $\nabla F(\mathbf{x}_t^i)$.

Clearly, different choices of the gradient estimator $\mathbf{v}_t^i$ lead to different instances of the `ProxGT` framework. Many local gradient estimation schemes are applicable here, such as the minibatch stochastic approximation [166, 177] and various variance reduction schemes, e.g., [49, 50, 59, 62]. As we explicitly show next, a suitable choice can be made in light of the underlying problem class and practical applications.

Algorithm 8 `ProxGT-SA` for Problem (6.1) with (6.2)

Ensure: Replace Line 2 in Algorithm 7 by the following for all $i \in \mathcal{V}$.

Require: b.

1: Obtain i.i.d samples $\{\boldsymbol{\xi}_{i,s}^t : s \in [b]\}$ for $\boldsymbol{\xi}_i$.

2: $\mathbf{v}_t^i := \frac{1}{b}\sum_{s=1}^{b} \nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t)$.

Algorithm 9 `ProxGT-SR-O` for Problem (6.1) with (6.2)

Ensure: Replace Line 2 in Algorithm 7 by the following for all $i \in \mathcal{V}$.

Require: B, b, q.

1: **if** $t \bmod q = 1$ **then**

2: Obtain i.i.d samples $\{\boldsymbol{\xi}_{i,s}^t : s \in [B]\}$ for $\boldsymbol{\xi}_i$.

3: Set $\mathbf{v}_t^i := \frac{1}{B}\sum_{s=1}^{B} \nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t)$.

4: **else**

5: Obtain i.i.d samples $\{\boldsymbol{\xi}_{i,s}^t : s \in [b]\}$ for $\boldsymbol{\xi}_i$.

6: $\mathbf{v}_t^i := \frac{1}{b}\sum_{s=1}^{b} \left(\nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t) - \nabla G_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_{i,s}^t)\right) + \mathbf{v}_{t-1}^i$.

7: **end if**

6.3.2 Instances of interest

In this section, we present several instances of `ProxGT` that are of particular interest for the population and empirical risk formulations considered in this chapter.

Expected risk minimization. A natural choice of the gradient estimator $\mathbf{v}_t^i$ in `ProxGT` is the minibatch stochastic approximation [177]. The resulting instance, called `ProxGT-SA`, is presented in Algorithm 8. An alternate approach is to construct the gradient estimator $\mathbf{v}_t^i$ in `ProxGT` via an online `SARAH` type recursive variance reduction scheme that effectively leverages the historical information to achieve faster convergence. When Assumption 6.2.5, the mean-squared smoothness, holds, the resulting algorithm `ProxGT-SR-O`, given in Algorithm 9, shows superior performance over Algorithm 8.

Empirical risk minimization. We now consider Problem (6.1) under the empirical risk (6.3). Although `ProxGT-SA` and `ProxGT-SR-O` developed in Section 6.3.2 remain applicable, the finite-sum structure of each f_i under (6.3) lends itself to faster stochastic variance reduction procedures [48–50]. In particular, we replace the periodic minibatch stochastic approximation step in `ProxGT-SR-O` by exact gradient computation. This corresponding implementation, named `ProxGT-SR-E`, is presented in Algorithm 10.

Algorithm 10 `ProxGT-SR-E` for Problem (6.1) with (6.3)

Ensure: Replace Line 2 in Algorithm 7 by the following for all $i \in \mathcal{V}$.

Require: b, q.

1: **if** $t \bmod q = 1$ **then**

2: Set $\mathbf{v}_t^i := \nabla f_i(\mathbf{x}_t^i)$.

3: **else**

4: Obtain i.i.d. samples $\{\boldsymbol{\xi}_{i,s}^t : s \in [b]\}$ for $\boldsymbol{\xi}_i$.

5: $\mathbf{v}_t^i := \frac{1}{b}\sum_{s=1}^b \big(\nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t) - \nabla G_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_{i,s}^t)\big) + \mathbf{v}_{t-1}^i$.

6: **end if**

6.4 Main results

In this section, we present the main convergence results of the proposed algorithms and highlight their implications. Throughout this section, we let Assumption 6.2.1, 6.2.2, and 6.2.3 hold without explicit statements. The iteration complexity of the `ProxGT` family is quantified in the following sense, while the gradient and communication complexities can be obtained accordingly.

Definition 6.4.1 (**Iteration Complexity**). *Consider the random vectors $\{\mathbf{x}_t^i\}$ generated by `ProxGT`. We say that `ProxGT` finds an ϵ-stationary point of Problem (6.1) in T iterations if*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}\big(\mathbf{x}_t^i\big)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \le \epsilon^2, \tag{6.12}$$

where $\overline{\mathbf{x}}_t := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_t^i$ and the gradient mapping $\mathbf{s}(\cdot)$ is defined in (6.7).

In view of Definition 6.2.1, if (6.12) holds true and we select the output, say $\widehat{\mathbf{x}}$, of `ProxGT` uniformly at random from $\{\mathbf{x}_t^i : t \in [T], i \in \mathcal{V}\}$, then $\mathbb{E}[\|\mathbf{s}(\widehat{\mathbf{x}})\|] \le \epsilon$, i.e., $\widehat{\mathbf{x}}$ is an ϵ-stationary solution for Problem (6.1).

6.4.1 Gradient and communication complexity

For ease of presentation, we define

$$\Delta := \Psi(\overline{\mathbf{x}}_1) - \underline{\Psi} \qquad \text{and} \qquad \zeta^2 := \tfrac{1}{n}\textstyle\sum_{i=1}^n \big\|\nabla f_i(\overline{\mathbf{x}}_1)\big\|^2. \tag{6.13}$$

Note that $\mathcal{O}(\cdot)$ in this section only hides universal constants that are not related to the problem parameters.

Theorem 6.4.1 (Convergence of `ProxGT-SA`). *Consider Problem (6.1) under the expected risk (6.2) and let Assumption 6.2.4 hold. Set $K \asymp \frac{\log(n\zeta)}{1-\lambda_*}$, $\alpha \asymp \frac{1}{L}$, $b \asymp \frac{\nu^2}{n\epsilon^2}$ in `ProxGT-SA`. Then `ProxGT-SA` finds an ϵ-stationary solution in $\mathcal{O}\big(\frac{L\Delta}{\epsilon^2}\big)$ iterations, leading to*

$$\mathcal{O}\left(\frac{L\Delta\nu^2}{n\epsilon^4}\right)$$

stochastic gradient samples at each node and

$$\mathcal{O}\left(\frac{L\Delta}{\epsilon^2}\cdot\frac{\log(n\zeta)}{1-\lambda_*}\right)$$

rounds of communication over the network.

In view of Theorem 6.4.1, `ProxGT-SA` achieves a *topology-independent* gradient complexity at each node that exhibits linear speedup against the centralized *optimal* minibatch stochastic proximal gradient method [149, 177] executed on a single node. To the best of our knowledge, this is the first time that such gradient complexity result is established in the literature of the general decentralized non-convex non-smooth composite expected risk minimization problems.

Theorem 6.4.2 (Convergence of `ProxGT-SR-O`)**.** *Consider Problem (6.1) under the expected risk (6.2) and let Assumption 6.2.4 and 6.2.5 hold. Set* $K \asymp \frac{\log(n\zeta)}{1-\lambda_*}$, $\alpha \asymp \frac{1}{L}$, $q \asymp \frac{\nu}{\epsilon}$, $b \asymp \frac{\nu}{n\epsilon}$, $B \asymp \frac{\nu^2}{n\epsilon^2}$ *in* `ProxGT-SR-O`*. Then* `ProxGT-SR-O` *finds an* ϵ*-stationary solution in* $\mathcal{O}\big(\frac{L\Delta}{\epsilon^2}+\frac{\nu}{\epsilon}\big)$ *iterations, leading to*

$$\mathcal{O}\left(\frac{L\Delta\nu}{n\epsilon^3}+\frac{\nu^2}{n\epsilon^2}\right)$$

stochastic gradient samples at each node and

$$\mathcal{O}\left(\left(\frac{L\Delta}{\epsilon^2}+\frac{\nu}{\epsilon}\right)\cdot\frac{\log(n\zeta)}{1-\lambda_*}\right)$$

rounds of communication over the network.

Theorem 6.4.2 shows that `ProxGT-SR-O` attains a *topology-independent* gradient complexity at each node that exhibits linear speedup compared to the centralized *optimal* proximal online variance reduction methods [48, 50, 149] implemented on a single node. To the best of our knowledge, this is the first such gradient complexity result in the literature of the decentralized non-convex non-smooth composite expected risk minimization problems with mean-squared smoothness.

For the special case $h = 0$, `ProxGT-SR-O` also improves the state-of-the-art gradient complexity result given by `GT-HSGD` [178] for smooth problems in the following sense. `GT-HSGD` achieves the optimal gradient complexity in the regime where the error tolerance ϵ of the problem is small enough, i.e., $\epsilon \lesssim (1-\lambda_*)^3 n^{-1}$. `ProxGT-SR-O` removes this regime restriction by performing $K \asymp \frac{\log(n\zeta)}{1-\lambda_*}$ rounds of consensus update per iteration.

Theorem 6.4.3 (Convergence of `ProxGT-SR-E`)**.** *Consider Problem (6.1) under the empirical risk (6.3) and let Assumption 6.2.4 and 6.2.5 hold. Set* $K \asymp \frac{\log\zeta}{1-\lambda_*}$, $\alpha \asymp \frac{1}{L}$, $q \asymp \sqrt{nm}$, $b \asymp \max\big\{\sqrt{\frac{m}{n}}, 1\big\}$ *in* `ProxGT-SR-E`*. Then* `ProxGT-SR-E` *finds an* ϵ*-stationary solution in* $\mathcal{O}\big(\frac{L\Delta}{\epsilon^2}+\sqrt{nm}\big)$ *iterations, leading to*

$$\mathcal{O}\left(\frac{L\Delta}{\epsilon^2}\max\left\{\sqrt{\frac{m}{n}},1\right\}+\max\left\{m,\sqrt{nm}\right\}\right)$$

stochastic gradient samples at each node and

$$\mathcal{O}\left(\left(\frac{L\Delta}{\epsilon^2}+\sqrt{nm}\right)\cdot\frac{\log\zeta}{1-\lambda_*}\right)$$

rounds of communication over the network.

Theorem 6.4.3 indicates that under a moderate big-data condition $m \gtrsim n$, `ProxGT-SR-E` achieves a *topology-independent* gradient complexity of $\mathcal{O}\left(\frac{L\Delta}{\epsilon^2}\sqrt{\frac{m}{n}}+m\right)$ at each node, leading to a linear speedup compared to the centralized *optimal* proximal finite-sum variance reduction methods [48, 50] implemented on a single node. To our knowledge, this is the first such gradient complexity result for decentralized non-convex non-smooth composite empirical risk minimization.

For the special case $h = 0$, `ProxGT-SR-E` also improves the state-of-the-art gradient complexity result achieved by `GT-SARAH` [141] for smooth problems in the following sense. The gradient complexity of `GT-SARAH` is optimal in the regime that the local sample size is large enough, i.e., $m \gtrsim n(1-\lambda_*)^{-6}$. `ProxGT-SR-E` improves this regime to $m \gtrsim n$ by performing $K \asymp \frac{\log\zeta}{1-\lambda_*}$ rounds of consensus updates per iteration.

Remark 6.4.1. In Theorem 6.4.1, 6.4.2, and 6.4.3, we set the number of communication rounds per iteration in `ProxGT` as a nontrivial constant $K > 1$ to obtain satisfactory gradient and communication complexities for `Prox-SA`, `Prox-SR-O`, and `Prox-SR-E` respectively. The complexity results of these algorithms for the case $K = 1$ can be obtained by slightly modifying the proofs of Theorem 6.4.1, 6.4.2, and 6.4.3 given in the appendix.

6.4.2 Improving communication complexity via accelerated consensus

It is possible to employ accelerated consensus algorithms, e.g., [77, 181], to implement the multiple consensus step $\mathbf{W}^K$ in `ProxGT` to achieve improved communication complexities when the network is undirected. The basic intuition is that the standard consensus algorithm $\mathbf{x}_{t+1} = \mathbf{W}\mathbf{x}_t$ returns an δ-average of the initial states in $\mathcal{O}\big(\frac{1}{1-\lambda_*}\log\frac{1}{\delta}\big)$ rounds of communication, while the accelerated consensus methods only take $\mathcal{O}\big(\frac{1}{\sqrt{1-\lambda_*}}\log\frac{1}{\delta}\big)$ rounds of communication.

In particular, we can replace $\mathbf{W}^K$ by a Chebyshev type polynomial of $\mathbf{W}$; see, e.g., [77, 142], for details. In this case, the communication complexity of `ProxGT-SA` stated in Theorem 6.4.1 improves to

$$\mathcal{O}\left(\frac{L\Delta}{\epsilon^2}\cdot\frac{\log(n\zeta)}{\sqrt{1-\lambda_*}}\right), \tag{6.14}$$

and the communication complexity of `ProxGT-SR-O` stated in Theorem 6.4.2 improves to

$$\mathcal{O}\left(\left(\frac{L\Delta}{\epsilon^2}+\frac{\nu}{\epsilon}\right)\cdot\frac{\log(n\zeta)}{\sqrt{1-\lambda_*}}\right), \tag{6.15}$$

Table 6.2: Datasets used in numerical experiments, available at https://www.openml.org/.

Dataset	**train** (nm)	**dimension** (p)
nomao	30,000	119
w8a	60,000	300
creditcard	100,000	29

and the communication complexity of `ProxGT-SR-E` stated in Theorem 6.4.3 improves to

$$\mathcal{O}\left(\left(\frac{L\Delta}{\epsilon^2}+\sqrt{nm}\right)\cdot\frac{\log\zeta}{\sqrt{1-\lambda_*}}\right), \tag{6.16}$$

while their sample complexities remain the same; we omit the detailed calculations here for conciseness. We note that the communication complexity of `ProxGT-SA` in (6.14) attains the lower bound provided in [179] for problems without the mean-squared smoothness and is hence optimal. Moreover, the communication complexity of `ProxGT-SR-O` and `ProxGT-SR-E` in (6.15) and (6.16) outperforms the respective state-of-the-art variance-reduced methods [141, 178] for smooth problems with mean-squared smoothness, with the help of multi-round accelerated consensus and proper mini-batches per iteration.

6.5 Numerical experiments

In this section, we numerically demonstrate the performance of the proposed `ProxGT` framework with the help of a sparse non-convex linear model [146] for decentralized binary classification problems.

Setup. In view of Problem (6.1), each local risk f_i and the convex non-smooth regularizer h take the following form:

$$f_i(\mathbf{x}) := \frac{1}{m}\sum_{j=1}^{m}\ell\left(b_{i,j}\cdot\mathbf{a}_{i,j}^\top\mathbf{x}\right) \quad \text{and} \quad h(\mathbf{x}) := c\|\mathbf{x}\|_1,$$

where $\ell:\mathbb{R}\to\mathbb{R}$ is given by

$$\ell(u) := \left(1-\frac{1}{1+\exp(-u)}\right)^2.$$

Here, $\mathbf{a}_{i,j}\in\mathbb{R}^p$ is the j-th feature vector at the i-th node, while $b_{i,j}\in\{-1,+1\}$ is the label for $\mathbf{a}_{i,j}$. It can be verified that ℓ is $\frac{4}{3}$-smooth and non-convex. We normalize each feature vector such that $\|\mathbf{a}_{i,j}\|=1, \forall i,j$, so that each f_i satisfies the mean-squared smoothness property with parameter $\frac{4}{3}$. We set $c=10^{-3}$ across all experiments, resulting in a relatively sparse solution, while the underlying network is an undirected geometric graph with 100 nodes. The doubly stochastic network weight matrix $\mathbf{W}_*$ is formulated by the lazy Metroplis rule [27], leading to a spectral gap $(1-\lambda_*)\approx 0.05$. A summary of the datasets used in the experiments is provided in Table 6.2.

Algorithms. Under the above formulation, we compare the `ProxGT-SA` and `ProxGT-SR-O` algorithms proposed in this chapter to `SPPDM` [176], the state-of-the-art decentralized stochastic proximal gradient

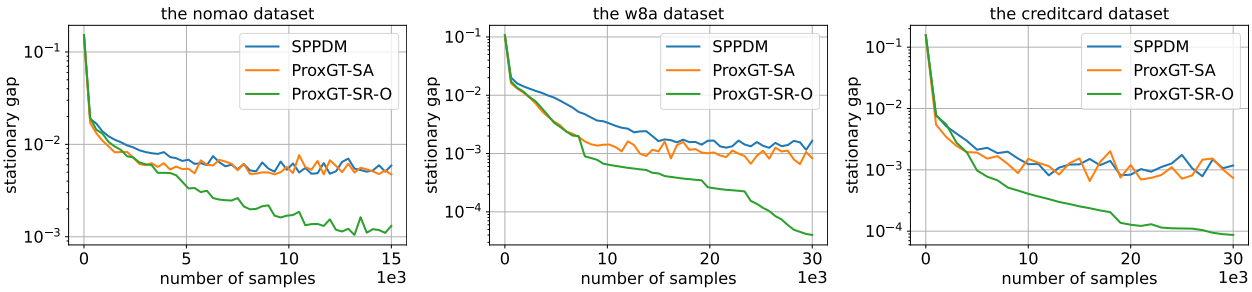

Figure 6.1: Sample efficiency comparison of `ProxGT-SA`, `ProxGT-SA-O`, and `SPPDM` on the nomao, w8a, and creditcard datasets over an undirected geometric graph with 100 nodes.

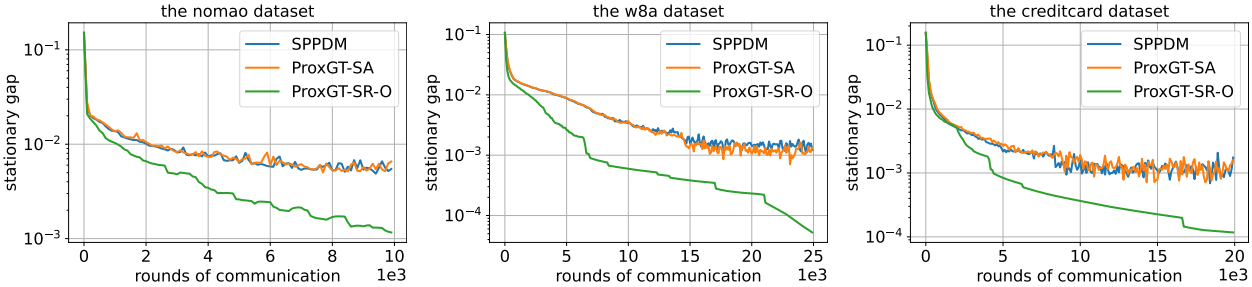

Figure 6.2: Communication efficiency comparison of `ProxGT-SA`, `ProxGT-SA-O`, and `SPPDM` on the nomao, w8a, and creditcard datasets over an undirected geometric graph with 100 nodes.

method. We do not consider `ProxGT-SR-E` here since it can be viewed as a special case of `ProxGT-SR-O` from a practical implementation viewpoint. We measure the performance of the algorithms in terms of the stationary gap $\|\mathbf{s}(\overline{\mathbf{x}})\|$, given in Definition 6.2.1, versus number of samples and rounds of communication, where $\overline{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ with $\mathbf{x}_i$ being the model at node i. For the sake of fair comparison, we set $K = 1$ in `ProxGT` and tune the algorithmic parameters by grid search for `ProxGT-SA`, `ProxGT-SR-O`, and `SPPDM`.

Observations. The experimental results are shown in Fig. 6.1 and 6.2. We observe that `ProxGT-SA` is more sample-efficient than `SPPDM` with a similar level of communication efficiency. Notably, `ProxGT-SR-O` significantly outperforms `ProxGT-SA` and `SPPDM` in terms of both gradient and communication efficiency, demonstrating the benefit of variance reduction.

Remark 6.5.1. It is worth mentioning that `SPPDM` is significantly more difficult to tune than the `ProxGT` family since the former has 7 algorithmic parameters to be optimized. Moreover, we emphasize that `SPPDM` is only provably applicable when the epigraph of h is a *polyhedral* set [176], while `ProxGT` does not have this restriction.

6.6 Outline of the convergence analysis

In this section, we describe a unified analysis for the proposed `Prox-GT` framework. Throughout the rest of the chapter, we let Assumption 6.2.1, 6.2.2, and 6.2.3 hold without explicit statements.

6.6.1 Preliminaries

We start by introducing some additional notations for Algorithm 7, 8, 9, and 10. We find it convenient to abstract the local proximal descent step by a *stochastic gradient mapping*: $\forall t \geq 1$ and $i \in \mathcal{V}$,

$$\mathbf{g}_t^i := \frac{1}{\alpha}\big(\mathbf{x}_t^i - \mathbf{z}_{t+1}^i\big). \tag{6.17}$$

For all $t \geq 1$, we let

$$\mathbf{g}_t := \begin{bmatrix} \mathbf{g}_t^1 \\ \vdots \\ \mathbf{g}_t^n \end{bmatrix}, \qquad \nabla \mathbf{f}(\mathbf{x}_t) := \begin{bmatrix} \nabla f_1(\mathbf{x}_t^1) \\ \vdots \\ \nabla f_n(\mathbf{x}_t^n) \end{bmatrix},$$

and define the following network mean states:

$$\begin{aligned} \overline{\mathbf{x}}_t &:= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_t^i, \qquad \overline{\mathbf{y}}_t := \frac{1}{n}\sum_{i=1}^n \mathbf{y}_t^i, \qquad \overline{\mathbf{z}}_t := \frac{1}{n}\sum_{i=1}^n \mathbf{z}_t^i, \\ \overline{\mathbf{v}}_t &:= \frac{1}{n}\sum_{i=1}^n \mathbf{v}_t^i, \qquad \overline{\mathbf{g}}_t := \frac{1}{n}\sum_{i=1}^n \mathbf{g}_t^i, \qquad \overline{\nabla \mathbf{f}}(\mathbf{x}_t) := \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_t^i). \end{aligned}$$

In addition, we define the exact averaging matrix

$$\mathbf{J} := \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right) \otimes \mathbf{I}_p.$$

Averaging (6.17) over i from 1 to n gives: $\forall t \geq 1$,

$$\overline{\mathbf{z}}_{t+1} = \overline{\mathbf{x}}_t - \alpha\overline{\mathbf{g}}_t. \tag{6.18}$$

We multiply $\frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)$ to the $\mathbf{x}$-update of Algorithm 7 to obtain: $\forall t \geq 1$,

$$\overline{\mathbf{x}}_{t+1} = \overline{\mathbf{z}}_{t+1}. \tag{6.19}$$

Combining (6.18) and (6.19) yields: $\forall t \geq 1$,

$$\overline{\mathbf{x}}_{t+1} = \overline{\mathbf{x}}_t - \alpha\overline{\mathbf{g}}_t. \tag{6.20}$$

Throughout the analysis, we fix arbitrary $K \geq 1$ and denote

$$\lambda := \lambda_*^K. \tag{6.21}$$

6.6.2 Basic facts

This section presents several basic facts that are used frequently in our analysis. We make use of a well-known non-expansiveness result for proximal mappings.

Lemma 6.6.1. [10] *Let $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed, and convex function. Then we have the following:*

$$\|\mathrm{prox}_h(\mathbf{x}) - \mathrm{prox}_h(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p.$$

For ease of reference, we give a trivial accumulation formula for scalar sequences with contraction.

Lemma 6.6.2. *Let $\{a_t\}$ and $\{b_t\}$ be scalar sequences and $0 < q < 1$, such that*

$$a_{t+1} \leq q a_t + b_t, \qquad \forall t \geq 1.$$

Then for all $T \geq 2$, we have

$$\sum_{t=1}^{T} a_t \leq \frac{1}{1-q} a_1 + \frac{1}{1-q} \sum_{t=1}^{T-1} b_t$$

and

$$\sum_{t=2}^{T+1} a_t \leq \frac{1}{1-q} a_2 + \frac{1}{1-q} \sum_{t=2}^{T} b_t.$$

Proof. The proof follows from standard arguments of convolution sums and we omit the details. □

The following lemma concerns the interplay between $\mathbf{W}$ and $\mathbf{J}$. We provide its proof for completeness.

Lemma 6.6.3. *The following statements hold for all $K \geq 1$.*

(a) $\mathbf{W}^K \mathbf{J} = \mathbf{J}\mathbf{W}^K = \mathbf{J}$.

(b) $\left\|\mathbf{W}^K - \mathbf{J}\right\| = \lambda_*^K$.

(c) $\left\|\mathbf{W}^K \mathbf{x} - \mathbf{J}\mathbf{x}\right\| \leq \lambda_*^K \left\|\mathbf{x} - \mathbf{J}\mathbf{x}\right\|, \forall \mathbf{x} \in \mathbb{R}^{np}$.

Proof. Since $\mathbf{W}_*$ is doubly stochastic, we have

$$\mathbf{W}\mathbf{J} = \mathbf{J}\mathbf{W} = \mathbf{J}, \tag{6.22}$$

which leads to part (a) by induction. Part (b) follows from

$$\left\|\mathbf{W}^K - \mathbf{J}\right\| = \left\|(\mathbf{W} - \mathbf{J})^K\right\| = \lambda_*^K,$$

where the first equality uses (6.22) and the second equality uses the definition of the spectral norm of a matrix. Finally, part (c) is due to

$$\left\|\mathbf{W}^K \mathbf{x} - \mathbf{J}\mathbf{x}\right\| = \left\|(\mathbf{W}^K - \mathbf{J})(\mathbf{x} - \mathbf{J}\mathbf{x})\right\| \leq \left\|\mathbf{W}^K - \mathbf{J}\right\| \left\|\mathbf{x} - \mathbf{J}\mathbf{x}\right\| = \lambda_*^K \left\|\mathbf{x} - \mathbf{J}\mathbf{x}\right\|,$$

where the first equality uses part (a) and $\mathbf{J}^2 = \mathbf{J}$, and the last equality uses part (b). □

Finally, we present a simple yet useful decomposition inequality.

Lemma 6.6.4. *Consider the iterates generated by Algorithm 7. Then we have:* $\forall T \geq 2$,

$$\sum_{t=2}^{T} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \leq 6 \sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + 3n\alpha^2 \sum_{t=1}^{T-1} \|\overline{\mathbf{g}}_t\|^2,$$

Proof. We note that $\forall t \geq 2$,

$$\begin{aligned}
\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 &= \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t + \mathbf{J}\mathbf{x}_t - \mathbf{J}\mathbf{x}_{t-1} + \mathbf{J}\mathbf{x}_{t-1} - \mathbf{x}_{t-1}\right\|^2 \\
&\leq 3\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + 3n\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_{t-1}\|^2 + 3\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2, \\
&\leq 3\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + 3n\alpha^2\|\overline{\mathbf{g}}_{t-1}\|^2 + 3\|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2, \qquad (6.23)
\end{aligned}$$

where the second line uses (6.20). Summing up (6.23) gives

$$\begin{aligned}
\sum_{t=2}^{T} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 &\leq 3 \sum_{t=2}^{T} \left(\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + \|\mathbf{x}_{t-1} - \mathbf{J}\mathbf{x}_{t-1}\|^2\right) + 3n\alpha^2 \sum_{t=2}^{T} \|\overline{\mathbf{g}}_{t-1}\|^2 \\
&\leq 6 \sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + +3n\alpha^2 \sum_{t=2}^{T} \|\overline{\mathbf{g}}_{t-1}\|^2
\end{aligned}$$

which finishes the proof. □

6.6.3 Descent inequality and error bounds

We first establish a key descent inequality in terms of the value of the global composite objective function Ψ. This result plays a central role in our analysis.

Lemma 6.6.5 (Descent). *Consider the iterates generated by Algorithm 7. If* $0 < \alpha \leq \frac{1}{8L}$, *then we have:* $\forall t \geq 1$,

$$\begin{aligned}
\frac{1}{n} \sum_{t=1}^{T} \left(\sum_{i=1}^{n} \left\|\mathbf{s}(\mathbf{x}_t^i)\right\|^2 + L^2 \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 \right) \leq \frac{8\Delta}{\alpha} - \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \left\|\mathbf{g}_t^i\right\|^2 + 76 \sum_{t=1}^{T} \left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 \\
+ \frac{6}{\alpha^2 n} \sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2 + \frac{10}{n} \sum_{t=2}^{T+1} \|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2.
\end{aligned}$$

Proof. See Section 6.7.1. □

In light of Lemma 6.6.5, our analysis approach is to show that the accumulated descent effect of the stochastic gradient mappings $\sum_{i=1}^{n} \|\mathbf{g}_t^i\|$ dominates the accumulated consensus, variance, and gradient tracking errors up to constant factors. To this aim, we establish useful error bounds for different algorithms. The following one is a consequence of the non-expansiveness of the proximal operator.

Lemma 6.6.6 (**Consensus**). *Consider the iterates generated by Algorithm 7. We have:* $\forall t \geq 1$,

$$\sum_{t=1}^{T} \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 \leq \frac{4\lambda^2\alpha^2}{(1-\lambda^2)^2} \sum_{t=2}^{T} \left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2.$$

Proof. See Section 6.7.2. □

Remark 6.6.1. It is worth noting that Lemma 6.6.5 and 6.6.6 do not use any properties of the gradient estimator $\mathbf{v}_t$. Therefore they may be of independent interest and used in other decentralized stochastic proximal gradient type methods for non-convex composite problems.

The next lemma establishes variance bounds for different algorithms.

Lemma 6.6.7 (**Variance**). *The following statements hold.*

(a) Let Assumption 6.2.4 hold and consider the iterates generated by Algorithm 8. Then we have: $\forall t \geq 1$,

$$\mathbb{E}\left[\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2\right] \leq \frac{\nu^2}{nb}.$$

(b) Let Assumption 6.2.4 and 6.2.5 hold. Consider the iterates generated by Algorithm 9. Suppose that $T = Rq$ *for some* $R \in \mathbb{Z}^+$. *Then we have:* $\forall T \geq q$,

$$\sum_{t=1}^{T} \mathbb{E}\left[\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2\right] \leq \frac{6L^2 q}{n^2 b} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + \frac{qL^2\alpha^2}{nb} \sum_{t=1}^{T-1} \mathbb{E}\left[\left\|\overline{\mathbf{g}}_t\right\|^2\right] + \frac{T\nu^2}{nB}.$$

(c) Let Assumption 6.2.5 hold. Consider the iterates generated by Algorithm 10. Suppose that $T = Rq$ *for some* $R \in \mathbb{Z}^+$. *Then we have:* $\forall T \geq q$,

$$\sum_{t=1}^{T} \mathbb{E}\left[\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2\right] \leq \frac{6L^2 q}{n^2 b} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + \frac{qL^2\alpha^2}{nb} \sum_{t=1}^{T-1} \mathbb{E}\left[\left\|\overline{\mathbf{g}}_t\right\|^2\right].$$

Proof. See Section 6.7.3. □

Finally, we give tracking error bounds for different algorithms in the following lemma.

Lemma 6.6.8 (**Tracking**). *The following statements hold.*

(a) Let Assumption 6.2.4 hold and consider the iterates generated by Algorithm 8. Then we have: $\forall T \geq 2$,

$$\sum_{t=2}^{T+1} \mathbb{E}\left[\left\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\right\|^2\right] \leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{12\lambda^2 n\alpha^2 L^2}{(1-\lambda^2)^2} \sum_{t=1}^{T-1} \mathbb{E}\left[\left\|\overline{\mathbf{g}}_t\right\|^2\right] + \frac{24\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right] + \frac{4T(2\lambda^2 n + 1)\nu^2}{b(1-\lambda^2)}.$$

(b) Let Assumption 6.2.4 and 6.2.5 hold. Consider the iterates generated by Algorithm 9. Let $T = Rq$ for some $R \in \mathbb{Z}^+$ and $R \geq 2$. Then we have:

$$\sum_{t=2}^{T+1} \mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right] \leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{96\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] + \frac{48\lambda^2 n\alpha^2 L^2}{(1-\lambda^2)^2} \sum_{t=1}^{T-1} \mathbb{E}\left[\|\overline{\mathbf{g}}_t\|^2\right] + \frac{14\lambda^2 T n\nu^2}{(1-\lambda^2)^2 Bq}.$$

(c) Let Assumption 6.2.5 hold. Consider the iterates generated by Algorithm 10. Let $T = Rq$ for some $R \in \mathbb{Z}^+$ and $R \geq 2$. Then we have:

$$\sum_{t=2}^{T+1} \mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right] \leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{96\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\right] + \frac{48\lambda^2 n\alpha^2 L^2}{(1-\lambda^2)^2} \sum_{t=1}^{T-1} \mathbb{E}\left[\|\overline{\mathbf{g}}_t\|^2\right].$$

Proof. See Section 6.7.4. □

6.6.4 Proofs of the main theorems

We first use the consensus error bound in Lemma 6.6.6 to refine the descent inequality in Lemma 6.6.5.

Proposition 6.6.1. *Consider the iterates generated by Algorithm 7. If $0 < \alpha \leq \frac{1}{8L}$, then we have: $\forall t \geq 1$,*

$$\frac{1}{n}\sum_{t=1}^{T}\left(\sum_{i=1}^{n}\left\|\mathbf{s}(\mathbf{x}_t^i)\right\|^2 + L^2\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right) \leq \frac{8\Delta}{\alpha} - \sum_{t=1}^{T}\|\overline{\mathbf{g}}_t\|^2 + 76\sum_{t=1}^{T}\left\|\overline{\mathbf{v}}_t - \overline{\nabla\mathbf{f}}(\mathbf{x}_t)\right\|^2 + \frac{34}{(1-\lambda^2)^2 n}\sum_{t=2}^{T+1}\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2.$$

Proof. This result follows by applying Lemma 6.6.6 to Lemma 6.6.5 and $\|\overline{\mathbf{g}}_t\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{g}_t^i\|^2$. □

6.6.4.1 Proof of Theorem 6.4.1

We apply Lemma 6.6.6 to Lemma 6.6.8(a) to obtain: $\forall T \geq 2$,

$$\left(1 - \frac{96\lambda^4\alpha^2 L^2}{(1-\lambda^2)^4}\right)\sum_{t=2}^{T+1}\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right] \leq \frac{2\lambda^2 n\zeta}{1-\lambda^2} + \frac{12\lambda^2 n\alpha^2 L^2}{(1-\lambda^2)^2}\sum_{t=1}^{T-1}\mathbb{E}\left[\|\overline{\mathbf{g}}_t\|^2\right] + \frac{4T(2\lambda^2 n+1)\nu^2}{b(1-\lambda^2)}. \tag{6.24}$$

If $0 < \alpha \leq \frac{(1-\lambda^2)^2}{14\lambda^2 L}$, then $1 - \frac{96\lambda^4\alpha^2 L^2}{(1-\lambda^2)^4} \geq \frac{1}{2}$ and hence (6.24) implies that $\forall T \geq 2$,

$$\sum_{t=2}^{T+1}\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\right] \leq \frac{4\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{24\lambda^2 n\alpha^2 L^2}{(1-\lambda^2)^2}\sum_{t=1}^{T-1}\mathbb{E}\left[\|\overline{\mathbf{g}}_t\|^2\right] + \frac{8T(2\lambda^2 n+1)\nu^2}{b(1-\lambda^2)}. \tag{6.25}$$

Plugging Lemma 6.6.7(a) and (6.25) into Proposition 6.6.1 gives: if $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{14\lambda^2}, \frac{1}{8}\right\}\frac{1}{L}$, then

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \\
&\leq \frac{8\Delta}{\alpha} - \sum_{t=1}^{T}\mathbb{E}\big[\|\overline{\mathbf{g}}_t\|^2\big] + \frac{76T\nu^2}{nb} + \frac{272T(2\lambda^2 n+1)\nu^2}{nb(1-\lambda^2)^3} + \frac{136\lambda^2\zeta^2}{(1-\lambda^2)^3} + \frac{816\lambda^2\alpha^2L^2}{(1-\lambda^2)^4}\sum_{t=1}^{T-1}\mathbb{E}\big[\|\overline{\mathbf{g}}_t\|^2\big] \\
&\leq \frac{8\Delta}{\alpha} - \left(1 - \frac{816\lambda^2\alpha^2L^2}{(1-\lambda^2)^4}\right)\sum_{t=1}^{T}\mathbb{E}\big[\|\overline{\mathbf{g}}_t\|^2\big] + \frac{136\lambda^2\zeta^2}{(1-\lambda^2)^3} + \frac{348T\nu^2}{nb(1-\lambda^2)^3} + \frac{544\lambda^2T\nu^2}{b(1-\lambda^2)^3}. \qquad (6.26)
\end{aligned}$$

From (6.26), we have: if $0 < \alpha \leq \min\left\{\frac{(1-\lambda^2)^2}{30\lambda}, \frac{1}{8}\right\}\frac{1}{L}$, then $\forall T \geq 2$,

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \leq \frac{8\Delta}{\alpha T} + \frac{136\lambda^2\zeta^2}{(1-\lambda^2)^3T} + \frac{348\nu^2}{nb(1-\lambda^2)^3} + \frac{544\lambda^2\nu^2}{b(1-\lambda^2)^3}. \qquad (6.27)$$

Recall from (6.21) that $\lambda := \lambda_*^K$ and we set

$$K \asymp \frac{\log(n\zeta)}{1-\lambda_*},$$

so that $\frac{1}{1-\lambda} = \mathcal{O}(1), \lambda\zeta = \mathcal{O}(1), \lambda n = \mathcal{O}(1)$. As a consequence, from (6.27) we have: if $0 < \alpha \lesssim \frac{1}{L}$, then

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \lesssim \frac{\Delta}{\alpha T} + \frac{\nu^2}{nb}. \qquad (6.28)$$

Finally, we observe that choosing

$$\alpha \asymp \frac{1}{L}, \qquad b \asymp \frac{\nu^2}{n\epsilon^2}, \qquad T \asymp \frac{L\Delta}{\epsilon^2}$$

in (6.28) gives $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{s}(\mathbf{x}_t^i)\|^2 + L^2\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\|^2\big] \lesssim \epsilon^2$. The ensuing complexity results follow from the fact that each iteration of Algorithm 8 incurs b stochastic gradient samples and K rounds of communication.

6.6.4.2 Proof of Theorem 6.4.2

Consider $T = Rq$ for some $R \in \mathbb{Z}^+$ and $R \geq 2$. Plugging Lemma 6.6.6 to Lemma 6.6.7(b) gives:

$$\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}_t - \overline{\nabla\mathbf{f}}(\mathbf{x}_t)\big\|^2\Big] \leq \frac{24\lambda^2\alpha^2L^2q}{(1-\lambda^2)^2n^2b}\sum_{t=2}^{T}\mathbb{E}\Big[\big\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\big\|^2\Big] + \frac{qL^2\alpha^2}{nb}\sum_{t=1}^{T-1}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big] + \frac{T\nu^2}{nB}.$$

In particular, if $0 < \alpha \leq \sqrt{\frac{nb}{24q}}\frac{1}{L}$, we have:

$$\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}_t - \overline{\nabla\mathbf{f}}(\mathbf{x}_t)\big\|^2\Big] \leq \frac{\lambda^2}{(1-\lambda^2)^2n}\sum_{t=2}^{T}\mathbb{E}\Big[\big\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\big\|^2\Big] + \frac{qL^2\alpha^2}{nb}\sum_{t=1}^{T-1}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big] + \frac{T\nu^2}{nB}. \qquad (6.29)$$

Applying (6.29) to Proposition 6.6.1 yields: if $0 < \alpha \leq \min\left\{\frac{1}{8}, \sqrt{\frac{nb}{24q}}\right\}\frac{1}{L}$, then

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] &\leq \frac{8\Delta}{\alpha} - \left(1 - \frac{76qL^2\alpha^2}{nb}\right)\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big] \\
&\quad + \frac{110}{(1-\lambda^2)^2n}\sum_{t=2}^{T+1}\mathbb{E}\Big[\big\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\big\|^2\Big] + \frac{76T\nu^2}{nB}.
\end{aligned}$$

In particular, if $0 < \alpha \leq \min\left\{\frac{1}{8}, \sqrt{\frac{nb}{152q}}\right\}\frac{1}{L}$, we have:

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \leq \frac{8\Delta}{\alpha} - \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big] + \frac{110}{(1-\lambda^2)^2 n}\sum_{t=2}^{T+1}\mathbb{E}\Big[\big\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\big\|^2\Big] + \frac{76T\nu^2}{nB}. \tag{6.30}$$

To proceed, we apply Lemma 6.6.6 to Lemma 6.6.8(b) to obtain:

$$\left(1 - \frac{384\lambda^4\alpha^2L^2}{(1-\lambda^2)^4}\right)\sum_{t=2}^{T+1}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] \leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{48\lambda^2 n\alpha^2L^2}{(1-\lambda^2)^2}\sum_{t=1}^{T-1}\mathbb{E}\big[\|\overline{\mathbf{g}}_t\|^2\big] + \frac{14\lambda^2 Tn\nu^2}{(1-\lambda^2)^2 Bq}. \tag{6.31}$$

If $0 < \alpha \leq \frac{(1-\lambda^2)^2}{28\lambda^2 L}$, (6.31) implies that

$$\sum_{t=2}^{T+1}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] \leq \frac{4\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{96\lambda^2 n\alpha^2L^2}{(1-\lambda^2)^2}\sum_{t=1}^{T-1}\mathbb{E}\big[\|\overline{\mathbf{g}}_t\|^2\big] + \frac{28\lambda^2 Tn\nu^2}{(1-\lambda^2)^2 Bq}. \tag{6.32}$$

Finally, plugging (6.32) to (6.30), we obtain: if $0 < \alpha \leq \min\left\{\frac{1}{8}, \sqrt{\frac{nb}{152q}}, \frac{(1-\lambda^2)^2}{28\lambda^2}\right\}\frac{1}{L}$, then

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \leq{}& \frac{8\Delta}{\alpha} + \frac{76T\nu^2}{nB} + \frac{440\lambda^2\zeta^2}{(1-\lambda^2)^3} + \frac{3080\lambda^2T\nu^2}{(1-\lambda^2)^4Bq} \\ &- \frac{1}{2}\left(1 - \frac{21120\lambda^2\alpha^2L^2}{(1-\lambda^2)^4}\right)\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big].\end{aligned}$$

Hence, if $0 < \alpha \leq \min\left\{\frac{1}{8}, \sqrt{\frac{nb}{152q}}, \frac{(1-\lambda^2)^2}{146\lambda^2}\right\}\frac{1}{L}$, then

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \leq \frac{8\Delta}{\alpha T} + \frac{76\nu^2}{nB} + \frac{440\lambda^2\zeta^2}{(1-\lambda^2)^3T} + \frac{3080\lambda^2\nu^2}{(1-\lambda^2)^4Bq}. \tag{6.33}$$

Let $\epsilon > 0$ be given. Recall from (6.21) that $\lambda := \lambda_*^K$ and we set

$$K \asymp \frac{\log(n\zeta)}{1-\lambda_*},$$

so that $\frac{1}{1-\lambda} = \mathcal{O}(1), \lambda\zeta = \mathcal{O}(1), \lambda n = \mathcal{O}(1)$; moreover, we let

$$q = nb \qquad \text{and} \qquad \alpha \asymp \frac{1}{L}.$$

As a consequence, we have from (6.33) that

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \lesssim \frac{L\Delta}{T} + \frac{\nu^2}{nB}. \tag{6.34}$$

In view of (6.34), we further choose

$$T \asymp \frac{L\Delta}{\epsilon^2} + q \qquad \text{and} \qquad B \asymp \frac{\nu^2}{n\epsilon^2}, \tag{6.35}$$

which lead to $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{s}(\mathbf{x}_t^i)\|^2+L^2\|\mathbf{x}_t^i-\overline{\mathbf{x}}_t\|^2\big]\lesssim\epsilon^2$. Since `ProxGT-SR-O` requires B samples every q iterations and b samples at each iteration, its total gradient complexity is bounded by

$$\mathcal{O}\left(T\left(b+\frac{B}{q}\right)\right). \tag{6.36}$$

Setting $b=B/q$, together with $q=nb$ stated above, gives

$$b\asymp\sqrt{\frac{B}{n}}=\frac{\nu}{n\epsilon},\qquad\text{and}\qquad q\asymp\frac{\nu}{\epsilon}. \tag{6.37}$$

Applying (6.35) and (6.37) to (6.36) concludes the ensuing gradient complexity and the corresponding communication complexity is given by TK.

6.6.4.3 Proof of Theorem 6.4.3

Consider $T=Rq$ for some $R\in\mathbb{Z}^+$ and $R\geq 2$. Plugging Lemma 6.6.6 to Lemma 6.6.7(c) gives:

$$\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}_t-\overline{\nabla\mathbf{f}}(\mathbf{x}_t)\big\|^2\Big]\leq\frac{24\lambda^2\alpha^2L^2q}{(1-\lambda^2)^2n^2b}\sum_{t=2}^{T}\mathbb{E}\Big[\big\|\mathbf{y}_t-\mathbf{J}\mathbf{y}_t\big\|^2\Big]+\frac{qL^2\alpha^2}{nb}\sum_{t=1}^{T-1}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big].$$

In particular, if $0<\alpha\leq\sqrt{\frac{nb}{24q}}\frac{1}{L}$, we have:

$$\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{v}}_t-\overline{\nabla\mathbf{f}}(\mathbf{x}_t)\big\|^2\Big]\leq\frac{\lambda^2}{(1-\lambda^2)^2n}\sum_{t=2}^{T}\mathbb{E}\Big[\big\|\mathbf{y}_t-\mathbf{J}\mathbf{y}_t\big\|^2\Big]+\frac{qL^2\alpha^2}{nb}\sum_{t=1}^{T-1}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big]. \tag{6.38}$$

Applying (6.38) to Proposition 6.6.1 yields: if $0<\alpha\leq\min\left\{\frac{1}{8},\sqrt{\frac{nb}{24q}}\right\}\frac{1}{L}$, then

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2+L^2\big\|\mathbf{x}_t^i-\overline{\mathbf{x}}_t\big\|^2\Big]\leq\frac{8\Delta}{\alpha}-\left(1-\frac{76qL^2\alpha^2}{nb}\right)\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big]\\+\frac{110}{(1-\lambda^2)^2n}\sum_{t=2}^{T+1}\mathbb{E}\Big[\big\|\mathbf{y}_t-\mathbf{J}\mathbf{y}_t\big\|^2\Big].\end{aligned}$$

In particular, if $0<\alpha\leq\min\left\{\frac{1}{8},\sqrt{\frac{nb}{152q}}\right\}\frac{1}{L}$, we have:

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2+L^2\big\|\mathbf{x}_t^i-\overline{\mathbf{x}}_t\big\|^2\Big]\leq\frac{8\Delta}{\alpha}-\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big]+\frac{110}{(1-\lambda^2)^2n}\sum_{t=2}^{T+1}\mathbb{E}\Big[\big\|\mathbf{y}_t-\mathbf{J}\mathbf{y}_t\big\|^2\Big]. \tag{6.39}$$

To proceed, we apply Lemma 6.6.6 to Lemma 6.6.8(b) to obtain:

$$\left(1-\frac{384\lambda^4\alpha^2L^2}{(1-\lambda^2)^4}\right)\sum_{t=2}^{T+1}\mathbb{E}\big[\|\mathbf{y}_t-\mathbf{J}\mathbf{y}_t\|^2\big]\leq\frac{2\lambda^2n\zeta^2}{1-\lambda^2}+\frac{48\lambda^2n\alpha^2L^2}{(1-\lambda^2)^2}\sum_{t=1}^{T-1}\mathbb{E}\big[\|\overline{\mathbf{g}}_t\|^2\big]. \tag{6.40}$$

If $0<\alpha\leq\frac{(1-\lambda^2)^2}{28\lambda^2L}$, (6.40) implies that

$$\sum_{t=2}^{T+1}\mathbb{E}\big[\|\mathbf{y}_t-\mathbf{J}\mathbf{y}_t\|^2\big]\leq\frac{4\lambda^2n\zeta^2}{1-\lambda^2}+\frac{96\lambda^2n\alpha^2L^2}{(1-\lambda^2)^2}\sum_{t=1}^{T-1}\mathbb{E}\big[\|\overline{\mathbf{g}}_t\|^2\big]. \tag{6.41}$$

Finally, plugging (6.41) to (6.39), we obtain: if $0 < \alpha \leq \min\left\{\frac{1}{8}, \sqrt{\frac{nb}{152q}}, \frac{(1-\lambda^2)^2}{28\lambda^2}\right\}\frac{1}{L}$, then

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \leq \frac{8\Delta}{\alpha} + \frac{440\lambda^2\zeta^2}{(1-\lambda^2)^3} - \frac{1}{2}\left(1 - \frac{21120\lambda^2\alpha^2L^2}{(1-\lambda^2)^4}\right)\sum_{t=1}^{T}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big].$$

Hence, if $0 < \alpha \leq \min\left\{\frac{1}{8}, \sqrt{\frac{nb}{152q}}, \frac{(1-\lambda^2)^2}{146\lambda^2}\right\}\frac{1}{L}$, then

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \leq \frac{8\Delta}{\alpha T} + \frac{440\lambda^2\zeta^2}{(1-\lambda^2)^3 T}. \tag{6.42}$$

Let $\epsilon > 0$ be given. Recall from (6.21) that $\lambda := \lambda_*^K$ and we set

$$K \asymp \frac{\log\zeta}{1-\lambda_*},$$

so that $\frac{1}{1-\lambda} = \mathcal{O}(1), \lambda\zeta = \mathcal{O}(1)$; moreover, we let

$$q = \sqrt{nm}, \qquad b = \max\left\{\sqrt{\frac{m}{n}}, 1\right\}, \qquad \alpha \asymp \frac{1}{L}. \tag{6.43}$$

As a consequence, we have from (6.42) that

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{s}(\mathbf{x}_t^i)\big\|^2 + L^2\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2\Big] \lesssim \frac{L\Delta}{T}. \tag{6.44}$$

In view of (6.44), we further choose

$$T \asymp \frac{L\Delta}{\epsilon^2} + q \tag{6.45}$$

which leads to $\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{s}(\mathbf{x}_t^i)\|^2 + L^2\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\|^2\big] \lesssim \epsilon^2$. The communication complexity is thus TK. Since `ProxGT-SR-E` requires m samples every q iterations and b samples at each iteration, its total gradient complexity is bounded by

$$\mathcal{O}\left(T\left(b + \frac{m}{q}\right)\right). \tag{6.46}$$

Plugging (6.43) and (6.45) into (6.46) concludes the ensuing gradient complexity.

6.7 Detailed proofs for lemmata in Section 6.6

6.7.1 Proof of Lemma 6.6.5

6.7.1.1 Step 1: Descent inequality for the convex part

First of all, we write the proximal descent step in Algorithm 7 in an equivalent form for analysis purposes. For all $t \geq 1$ and $i \in \mathcal{V}$, we observe that

$$\begin{aligned}\mathbf{z}_{t+1}^i = \mathrm{prox}_{\alpha h}\big(\mathbf{x}_t^i - \alpha \mathbf{y}_{t+1}^i\big) &= \mathrm{argmin}_{\mathbf{u}\in\mathbb{R}^p}\left\{\frac{1}{2}\big\|\mathbf{u} - (\mathbf{x}_t^i - \alpha\mathbf{y}_{t+1}^i)\big\|^2 + \alpha h(\mathbf{u})\right\} \\ &= \mathrm{argmin}_{\mathbf{u}\in\mathbb{R}^p}\left\{\frac{1}{2}\big\|\mathbf{u} - \mathbf{x}_t^i\big\|^2 + \big\langle \alpha\mathbf{y}_{t+1}^i, \mathbf{u} - \mathbf{x}_t^i\big\rangle + \frac{1}{2}\big\|\alpha\mathbf{y}_{t+1}^i\big\|^2 + \alpha h(\mathbf{u})\right\} \\ &= \mathrm{argmin}_{\mathbf{u}\in\mathbb{R}^p}\left\{\big\langle \mathbf{y}_{t+1}^i, \mathbf{u}\big\rangle + \frac{1}{2\alpha}\big\|\mathbf{u} - \mathbf{x}_t^i\big\|^2 + h(\mathbf{u})\right\}. \end{aligned} \tag{6.47}$$

In light of the optimality condition of the strongly convex optimization problem (6.47) and the sum rule of subdifferential calculus [10], for all $t \geq 1$ and $i \in \mathcal{V}$, there exists $h'(\mathbf{z}_{t+1}^i) \in \partial h(\mathbf{z}_{t+1}^i)$ such that

$$h'(\mathbf{z}_{t+1}^i) = -\mathbf{y}_{t+1}^i - \frac{1}{\alpha}\big(\mathbf{z}_{t+1}^i - \mathbf{x}_t^i\big). \tag{6.48}$$

By the subgradient inequality, we have: $\forall t \geq 1$, $\forall i \in \mathcal{V}$, and $\forall \mathbf{u} \in \mathbb{R}^p$,

$$h(\mathbf{u}) \geq h(\mathbf{z}_{t+1}^i) + \big\langle h'(\mathbf{z}_{t+1}^i), \mathbf{u} - \mathbf{z}_{t+1}^i\big\rangle,$$

which is the same as

$$h(\mathbf{z}_{t+1}^i) \leq h(\mathbf{u}) + \big\langle h'(\mathbf{z}_{t+1}^i), \mathbf{z}_{t+1}^i - \mathbf{u}\big\rangle. \tag{6.49}$$

Applying (6.48) to (6.49), we obtain: $\forall t \geq 1$, $\forall i \in \mathcal{V}$, and $\forall \mathbf{u} \in \mathbb{R}^p$,

$$h(\mathbf{z}_{t+1}^i) \leq h(\mathbf{u}) - \frac{1}{\alpha}\big\langle \mathbf{x}_t^i - \mathbf{z}_{t+1}^i, \mathbf{u} - \mathbf{z}_{t+1}^i\big\rangle - \big\langle \mathbf{y}_{t+1}^i, \mathbf{z}_{t+1}^i - \mathbf{u}\big\rangle. \tag{6.50}$$

We have the following algebraic identity: $\forall t \geq 1$, $\forall i \in \mathcal{V}$, and $\forall \mathbf{u} \in \mathbb{R}^p$,

$$\big\langle \mathbf{x}_t^i - \mathbf{z}_{t+1}^i, \mathbf{u} - \mathbf{z}_{t+1}^i\big\rangle = \frac{1}{2}\big\|\mathbf{u} - \mathbf{z}_{t+1}^i\big\|^2 + \frac{1}{2}\big\|\mathbf{x}_t^i - \mathbf{z}_{t+1}^i\big\|^2 - \frac{1}{2}\big\|\mathbf{x}_t^i - \mathbf{u}\big\|^2. \tag{6.51}$$

Applying (6.51) to (6.50), we obtain: $\forall t \geq 1$, $\forall i \in \mathcal{V}$, and $\forall \mathbf{u} \in \mathbb{R}^p$,

$$h(\mathbf{z}_{t+1}^i) \leq h(\mathbf{u}) - \frac{1}{2\alpha}\big\|\mathbf{u} - \mathbf{z}_{t+1}^i\big\|^2 - \frac{1}{2\alpha}\big\|\mathbf{x}_t^i - \mathbf{z}_{t+1}^i\big\|^2 + \frac{1}{2\alpha}\big\|\mathbf{x}_t^i - \mathbf{u}\big\|^2 - \big\langle \mathbf{y}_{t+1}^i, \mathbf{z}_{t+1}^i - \mathbf{u}\big\rangle. \tag{6.52}$$

Setting $\mathbf{u} := \overline{\mathbf{x}}_t$, we have: $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\begin{aligned} h(\mathbf{z}_{t+1}^i) &\leq h(\overline{\mathbf{x}}_t) - \frac{1}{2\alpha}\big\|\overline{\mathbf{x}}_t - \mathbf{z}_{t+1}^i\big\|^2 - \frac{1}{2\alpha}\big\|\mathbf{x}_t^i - \mathbf{z}_{t+1}^i\big\|^2 + \frac{1}{2\alpha}\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2 - \big\langle \mathbf{y}_{t+1}^i, \mathbf{z}_{t+1}^i - \overline{\mathbf{x}}_t\big\rangle \\ &= h(\overline{\mathbf{x}}_t) - \frac{1}{2\alpha}\big\|\overline{\mathbf{x}}_t - \mathbf{z}_{t+1}^i\big\|^2 - \frac{\alpha}{2}\big\|\mathbf{g}_t^i\big\|^2 + \frac{1}{2\alpha}\big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\big\|^2 - \big\langle \mathbf{y}_{t+1}^i - \overline{\mathbf{y}}_{t+1}, \mathbf{z}_{t+1}^i - \overline{\mathbf{x}}_t\big\rangle \\ &\quad - \big\langle \overline{\mathbf{y}}_{t+1}, \mathbf{z}_{t+1}^i - \overline{\mathbf{x}}_t\big\rangle, \end{aligned} \tag{6.53}$$

where the last line uses (6.17). For the second last term in (6.53), we have: $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\begin{aligned} -\left\langle \mathbf{y}_{t+1}^i - \overline{\mathbf{y}}_{t+1}, \mathbf{z}_{t+1}^i - \overline{\mathbf{x}}_t \right\rangle &\leq \|\mathbf{y}_{t+1}^i - \overline{\mathbf{y}}_{t+1}\| \|\mathbf{z}_{t+1}^i - \overline{\mathbf{x}}_t\| \\ &\leq \frac{\alpha}{2}\|\mathbf{y}_{t+1}^i - \overline{\mathbf{y}}_{t+1}\|^2 + \frac{1}{2\alpha}\|\mathbf{z}_{t+1}^i - \overline{\mathbf{x}}_t\|^2, \end{aligned} \tag{6.54}$$

where the first and the second line use the Cauchy-Schwarz and Young's inequality respectively. Plugging (6.54) into (6.53) gives: $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$h(\mathbf{z}_{t+1}^i) \leq h(\overline{\mathbf{x}}_t) - \frac{\alpha}{2}\left\|\mathbf{g}_t^i\right\|^2 + \frac{1}{2\alpha}\left\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t\right\|^2 + \frac{\alpha}{2}\|\mathbf{y}_{t+1}^i - \overline{\mathbf{y}}_{t+1}\|^2 - \left\langle \overline{\mathbf{y}}_{t+1}, \mathbf{z}_{t+1}^i - \overline{\mathbf{x}}_t \right\rangle. \tag{6.55}$$

We now average (6.55) over i from 1 to n to obtain: $\forall t \geq 1$,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n h(\mathbf{z}_{t+1}^i) &\leq h(\overline{\mathbf{x}}_t) - \frac{\alpha}{2n}\sum_{i=1}^n \left\|\mathbf{g}_t^i\right\|^2 + \frac{1}{2\alpha n}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \frac{\alpha}{2n}\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 - \left\langle \overline{\mathbf{y}}_{t+1}, \overline{\mathbf{z}}_{t+1} - \overline{\mathbf{x}}_t \right\rangle \\ &= h(\overline{\mathbf{x}}_t) - \frac{\alpha}{2n}\sum_{i=1}^n \left\|\mathbf{g}_t^i\right\|^2 + \frac{1}{2\alpha n}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \frac{\alpha}{2n}\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 + \alpha\left\langle \overline{\mathbf{y}}_{t+1}, \overline{\mathbf{g}}_t \right\rangle, \end{aligned} \tag{6.56}$$

where the second line follows from (6.18). In light of the convexity of h and Jensen's inequality, for all $t \geq 1$ we have that $h(\overline{\mathbf{z}}_{t+1}) \leq \frac{1}{n}\sum_{i=1}^n h(\mathbf{z}_{t+1}^i)$ and hence (6.56) implies

$$h(\overline{\mathbf{z}}_{t+1}) \leq h(\overline{\mathbf{x}}_t) - \frac{\alpha}{2n}\sum_{i=1}^n \left\|\mathbf{g}_t^i\right\|^2 + \frac{1}{2\alpha n}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \frac{\alpha}{2n}\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 + \alpha\left\langle \overline{\mathbf{y}}_{t+1}, \overline{\mathbf{g}}_t \right\rangle, \quad \forall t \geq 1. \tag{6.57}$$

In view of (6.19), we observe that (6.57) is the same as

$$h(\overline{\mathbf{x}}_{t+1}) \leq h(\overline{\mathbf{x}}_t) - \frac{\alpha}{2n}\sum_{i=1}^n \left\|\mathbf{g}_t^i\right\|^2 + \frac{1}{2\alpha n}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \frac{\alpha}{2n}\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 + \alpha\left\langle \overline{\mathbf{y}}_{t+1}, \overline{\mathbf{g}}_t \right\rangle, \quad \forall t \geq 1. \tag{6.58}$$

6.7.1.2 Step 2: Descent inequality for the non-convex part

Since F is L-smooth, we have the standard quadratic upper bound [10]:

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p. \tag{6.59}$$

Setting $\mathbf{y} = \overline{\mathbf{x}}_{t+1}$ and $\mathbf{x} = \overline{\mathbf{x}}_t$ in (6.59), we obtain: $\forall t \geq 1$,

$$\begin{aligned} F(\overline{\mathbf{x}}_{t+1}) &\leq F(\overline{\mathbf{x}}_t) + \langle \nabla F(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_{t+1} - \overline{\mathbf{x}}_t \rangle + \frac{L}{2}\|\overline{\mathbf{x}}_{t+1} - \overline{\mathbf{x}}_t\|^2 \\ &= F(\overline{\mathbf{x}}_t) - \alpha\langle \nabla F(\overline{\mathbf{x}}_t), \overline{\mathbf{g}}_t \rangle + \frac{L\alpha^2}{2}\|\overline{\mathbf{g}}_t\|^2, \end{aligned} \tag{6.60}$$

where the last line is due to (6.20).

6.7.1.3 Step 3: combining step 1 and step 2

Recall that $\Psi := F + h$. Summing up (6.60) and (6.58), we obtain: $\forall t \geq 1$,

$$\begin{aligned} \Psi(\overline{\mathbf{x}}_{t+1}) \leq \Psi(\overline{\mathbf{x}}_t) &- \frac{\alpha}{2n}\sum_{i=1}^n \left\|\mathbf{g}_t^i\right\|^2 + \frac{1}{2\alpha n}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \frac{\alpha}{2n}\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 \\ &+ \alpha\left\langle \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t), \overline{\mathbf{g}}_t \right\rangle + \frac{L\alpha^2}{2}\|\overline{\mathbf{g}}_t\|^2. \end{aligned} \tag{6.61}$$

By the Cauchy-Schwarz and Young's inequality, we have: $\forall \eta > 0$ and $\forall t \geq 1$,

$$\left\langle \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t), \overline{\mathbf{g}}_t \right\rangle \leq \frac{1}{2\eta} \left\| \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t) \right\|^2 + \frac{\eta}{2} \left\| \overline{\mathbf{g}}_t \right\|^2 \tag{6.62}$$

Applying (6.62) to (6.61), we obtain: $\forall \eta > 0$ and $\forall t \geq 1$,

$$\begin{aligned} \Psi(\overline{\mathbf{x}}_{t+1}) \leq \Psi(\overline{\mathbf{x}}_t) - \frac{\alpha}{2n} \sum_{i=1}^n \left\| \mathbf{g}_t^i \right\|^2 + \frac{1}{2\alpha n} \left\| \mathbf{x}_t - \mathbf{J}\mathbf{x}_t \right\|^2 + \frac{\alpha}{2n} \| \mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1} \|^2 \\ + \frac{\alpha}{2\eta} \left\| \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t) \right\|^2 + \frac{\eta\alpha + L\alpha^2}{2} \| \overline{\mathbf{g}}_t \|^2. \end{aligned} \tag{6.63}$$

6.7.1.4 Step 4: Refining error terms and telescoping sum

We first bound the difference between the local stochastic gradient mapping $\mathbf{g}_t^i$ defined in (6.17) and the exact gradient mapping $\mathbf{s}(\mathbf{x}_t^i)$ defined in (6.7). Observe that $\forall t \geq 1$ and $\forall i \in \mathcal{V}$,

$$\begin{aligned} \left\| \mathbf{g}_t^i - \mathbf{s}(\mathbf{x}_t^i) \right\|^2 &= \left\| \frac{1}{\alpha} \Big(\mathbf{x}_t^i - \mathrm{prox}_{\alpha h}\big(\mathbf{x}_t^i - \alpha \mathbf{y}_{t+1}^i \big) \Big) - \frac{1}{\alpha} \Big(\mathbf{x}_t^i - \mathrm{prox}_{\alpha h}\big(\mathbf{x}_t^i - \alpha \nabla F(\mathbf{x}_t^i) \big) \Big) \right\|^2 \\ &= \frac{1}{\alpha^2} \left\| \mathrm{prox}_{\alpha h}\big(\mathbf{x}_t^i - \alpha \mathbf{y}_{t+1}^i \big) - \mathrm{prox}_{\alpha h}\big(\mathbf{x}_t^i - \alpha \nabla F(\mathbf{x}_t^i) \big) \right\|^2 \\ &\leq \left\| \mathbf{y}_{t+1}^i - \nabla F(\mathbf{x}_t^i) \right\|^2 \\ &= \left\| \mathbf{y}_{t+1}^i - \overline{\mathbf{y}}_{t+1} + \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t) + \nabla F(\overline{\mathbf{x}}_t) - \nabla F(\mathbf{x}_t^i) \right\|^2 \\ &\leq 3 \left\| \mathbf{y}_{t+1}^i - \overline{\mathbf{y}}_{t+1} \right\|^2 + 3 \left\| \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t) \right\|^2 + 3L^2 \left\| \overline{\mathbf{x}}_t - \mathbf{x}_t^i \right\|^2, \end{aligned} \tag{6.64}$$

where the third line is due to Lemma 6.6.1 and the last line uses the L-smoothness of F. Observe that $\forall t \geq 1$,

$$\begin{aligned} -\left\| \mathbf{g}_t^i \right\|^2 &\leq -\frac{1}{2} \left\| \mathbf{s}(\mathbf{x}_t^i) \right\|^2 + \left\| \mathbf{g}_t^i - \mathbf{s}(\mathbf{x}_t^i) \right\|^2 \\ &\leq -\frac{1}{2} \left\| \mathbf{s}(\mathbf{x}_t^i) \right\|^2 + 3 \left\| \mathbf{y}_{t+1}^i - \overline{\mathbf{y}}_{t+1} \right\|^2 + 3 \left\| \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t) \right\|^2 + 3L^2 \left\| \overline{\mathbf{x}}_t - \mathbf{x}_t^i \right\|^2, \end{aligned} \tag{6.65}$$

where the first line is due to the standard triangular inequality and the second line uses (6.64). Averaging (6.65) over i from 1 to n gives: $\forall t \geq 1$,

$$-\frac{1}{n} \sum_{i=1}^n \left\| \mathbf{g}_t^i \right\|^2 \leq -\frac{1}{2n} \sum_{i=1}^n \left\| \mathbf{s}(\mathbf{x}_t^i) \right\|^2 + \frac{3}{n} \| \mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1} \|^2 + 3 \left\| \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t) \right\|^2 + \frac{3L^2}{n} \left\| \mathbf{x}_t - \mathbf{J}\mathbf{x}_t \right\|^2. \tag{6.66}$$

We now plug (6.66) into (6.63) to obtain: $\forall \eta > 0$ and $\forall t \geq 1$,

$$\begin{aligned} \Psi(\overline{\mathbf{x}}_{t+1}) &\leq \Psi(\overline{\mathbf{x}}_t) - \frac{\alpha}{4n} \sum_{i=1}^n \left\| \mathbf{g}_t^i \right\|^2 - \frac{\alpha}{8n} \sum_{i=1}^n \left\| \mathbf{s}(\mathbf{x}_t^i) \right\|^2 + \left(\frac{1}{2\alpha} + \frac{3\alpha L^2}{4} \right) \frac{1}{n} \left\| \mathbf{x}_t - \mathbf{J}\mathbf{x}_t \right\|^2 + \frac{5\alpha}{4n} \| \mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1} \|^2 \\ &\quad + \left(\frac{3}{2} + \frac{1}{\eta} \right) \frac{\alpha}{2} \left\| \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t) \right\|^2 + \frac{\eta\alpha + L\alpha^2}{2} \| \overline{\mathbf{g}}_t \|^2 \\ &\leq \Psi(\overline{\mathbf{x}}_t) - \frac{\alpha - 2\eta\alpha - 2L\alpha^2}{4n} \sum_{i=1}^n \left\| \mathbf{g}_t^i \right\|^2 - \frac{\alpha}{8n} \sum_{i=1}^n \left\| \mathbf{s}(\mathbf{x}_t^i) \right\|^2 + \left(\frac{1}{2\alpha} + \frac{3\alpha L^2}{4} \right) \frac{1}{n} \left\| \mathbf{x}_t - \mathbf{J}\mathbf{x}_t \right\|^2 \\ &\quad + \frac{5\alpha}{4n} \| \mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1} \|^2 + \left(\frac{3}{2} + \frac{1}{\eta} \right) \frac{\alpha}{2} \left\| \overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t) \right\|^2, \end{aligned} \tag{6.67}$$

where the last line is due to $\|\overline{\mathbf{g}}_t\|^2 \leq \frac{1}{n}\sum_{i=1}^n \|\mathbf{g}_t^i\|^2$. Setting $\eta = \frac{1}{8}$ and $0 < \alpha \leq \frac{1}{8L}$ in (6.67), we have: $\forall t \geq 1$,

$$\Psi(\overline{\mathbf{x}}_{t+1}) \leq \Psi(\overline{\mathbf{x}}_t) - \frac{\alpha}{8n}\sum_{i=1}^n \left\|\mathbf{g}_t^i\right\|^2 - \frac{\alpha}{8n}\sum_{i=1}^n \left\|\mathbf{s}(\mathbf{x}_t^i)\right\|^2 + \left(\frac{1}{2\alpha} + \frac{3\alpha L^2}{4}\right)\frac{1}{n}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \frac{5\alpha}{4n}\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 + \frac{19\alpha}{4}\left\|\overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t)\right\|^2. \tag{6.68}$$

Towards the last term in (6.68), observe that, $\forall t \geq 1$,

$$\begin{aligned}\left\|\overline{\mathbf{y}}_{t+1} - \nabla F(\overline{\mathbf{x}}_t)\right\|^2 &= \|\overline{\mathbf{v}}_t - \nabla F(\overline{\mathbf{x}}_t)\|^2 \\ &\leq 2\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + 2\left\|\overline{\nabla \mathbf{f}}(\mathbf{x}_t) - \nabla F(\overline{\mathbf{x}}_t)\right\|^2 \\ &\leq 2\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + 2L^2 n^{-1}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2,\end{aligned} \tag{6.69}$$

where the first line is due to (6.11) while the last line uses the L-smoothness of each f_i, i.e.,

$$\left\|\overline{\nabla \mathbf{f}}(\mathbf{x}_t) - \nabla F(\overline{\mathbf{x}}_t)\right\|^2 = \left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(\mathbf{x}_t^i) - \nabla f_i(\overline{\mathbf{x}}_t)\right)\right\|^2 \leq \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}_t^i) - \nabla f_i(\overline{\mathbf{x}}_t)\right\|^2 \leq \frac{L^2}{n}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2.$$

Plugging (6.69) into (6.68), we have: if $0 < \alpha \leq \frac{1}{8L}$, then $\forall t \geq 1$,

$$\Psi(\overline{\mathbf{x}}_{t+1}) \leq \Psi(\overline{\mathbf{x}}_t) - \frac{\alpha}{8n}\sum_{i=1}^n \left\|\mathbf{g}_t^i\right\|^2 - \frac{\alpha}{8n}\sum_{i=1}^n \left\|\mathbf{s}(\mathbf{x}_t^i)\right\|^2 + \left(\frac{1}{2\alpha} + \frac{41\alpha L^2}{4}\right)\frac{1}{n}\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \frac{5\alpha}{4n}\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 + \frac{19\alpha}{2}\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2. \tag{6.70}$$

Telescoping sum (6.70) over t from 1 to T, we have: if $0 < \alpha \leq \frac{1}{8L}$, then

$$\Psi(\overline{\mathbf{x}}_{T+1}) \leq \Psi(\overline{\mathbf{x}}_1) - \frac{\alpha}{8n}\sum_{t=1}^T\sum_{i=1}^n \left\|\mathbf{g}_t^i\right\|^2 - \frac{\alpha}{8n}\sum_{t=1}^T\sum_{i=1}^n \left\|\mathbf{s}(\mathbf{x}_t^i)\right\|^2 + \left(\frac{1}{2\alpha} + \frac{41\alpha L^2}{4}\right)\frac{1}{n}\sum_{t=1}^T \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \frac{5\alpha}{4n}\sum_{t=1}^T \|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 + \frac{19\alpha}{2}\sum_{t=1}^T \left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2. \tag{6.71}$$

With $\inf_{\mathbf{x}\in\mathbb{R}^p} \Psi(\mathbf{x}) \geq \underline{\Psi} > -\infty$ and minor rearrangement, (6.71) implies the following: if $0 < \alpha \leq \frac{1}{8L}$, then

$$\frac{1}{n}\sum_{t=1}^T \left(\sum_{i=1}^n \left\|\mathbf{s}(\mathbf{x}_t^i)\right\|^2 + L^2\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\right) \leq \frac{8(\Psi(\overline{\mathbf{x}}_1) - \underline{\Psi})}{\alpha} - \frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n \left\|\mathbf{g}_t^i\right\|^2 + 76\sum_{t=1}^T \left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 + \left(\frac{4}{\alpha^2} + 83L^2\right)\frac{1}{n}\sum_{t=1}^T \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 + \frac{10}{n}\sum_{t=2}^{T+1} \|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2,$$

which finishes the proof of Lemma 6.6.5 by $83L^2 \leq \frac{2}{\alpha^2}$.

6.7.2 Proof of Lemma 6.6.6

For ease of exposition, we define a block-wise proximal mapping for h:

$$\mathbf{prox}_{\alpha h}(\mathbf{c}) := \begin{bmatrix}\mathrm{prox}_{\alpha h}(\mathbf{c}_1) \\ \vdots \\ \mathrm{prox}_{\alpha h}(\mathbf{c}_n)\end{bmatrix} \in \mathbb{R}^{np}, \quad \text{where} \quad \mathbf{c} := \begin{bmatrix}\mathbf{c}_1 \\ \vdots \\ \mathbf{c}_n\end{bmatrix} \tag{6.72}$$

such that $\mathbf{c}_i \in \mathbb{R}^p, \forall i \in [n]$. In view of (6.72), the $\mathbf{x}$-update in Algorithm 7 can compactly be written as

$$\mathbf{x}_{t+1} = \mathbf{W}^K \mathbf{prox}_{\alpha h}(\mathbf{x}_t - \alpha \mathbf{y}_{t+1}), \qquad \forall t \geq 1. \tag{6.73}$$

We find the following quantity helpful: $\forall t \geq 1$,

$$\begin{aligned}
\big(\mathbf{W}^K - \mathbf{J}\big)\mathbf{prox}_{\alpha h}(\mathbf{J}\mathbf{x}_t - \alpha \mathbf{J}\mathbf{y}_{t+1}) &= \Big(\big(\mathbf{W}_*^K - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\big) \otimes \mathbf{I}_p\Big)\Big(\mathbf{1}_n \otimes \mathrm{prox}_{\alpha h}\big(\overline{\mathbf{x}}_t - \alpha \overline{\mathbf{y}}_{t+1}\big)\Big) \\
&= \Big(\big(\mathbf{W}_*^K - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\big)\mathbf{1}_n\Big) \otimes \mathrm{prox}_{\alpha h}\big(\overline{\mathbf{x}}_t - \alpha \overline{\mathbf{y}}_{t+1}\big) \\
&= \mathbf{0}_{np},
\end{aligned} \tag{6.74}$$

where the first line uses the definition of $\mathbf{W}$, $\mathbf{J}$, and $\mathbf{prox}_{\alpha h}$, and the last line is due to the doubly stochasticity of $\mathbf{W}_*$. We are now prepared to analyze the consensus error recursion in the following. For all $t \geq 1$, we have

$$\begin{aligned}
\big\|\mathbf{x}_{t+1} - \mathbf{J}\mathbf{x}_{t+1}\big\|^2 &= \Big\|\mathbf{W}^K \mathbf{prox}_{\alpha h}(\mathbf{x}_t - \alpha \mathbf{y}_{t+1}) - \mathbf{J}\mathbf{W}^K \mathbf{prox}_{\alpha h}(\mathbf{x}_t - \alpha \mathbf{y}_{t+1})\Big\|^2 \\
&= \Big\|\big(\mathbf{W}^K - \mathbf{J}\big)\mathbf{prox}_{\alpha h}(\mathbf{x}_t - \alpha \mathbf{y}_{t+1})\Big\|^2 \\
&= \Big\|\big(\mathbf{W}^K - \mathbf{J}\big)\Big(\mathbf{prox}_{\alpha h}(\mathbf{x}_t - \alpha \mathbf{y}_{t+1}) - \mathbf{prox}_{\alpha h}(\mathbf{J}\mathbf{x}_t - \alpha \mathbf{J}\mathbf{y}_{t+1})\Big)\Big\|^2 \\
&\leq \lambda^2 \Big\|\mathbf{prox}_{\alpha h}(\mathbf{x}_t - \alpha \mathbf{y}_{t+1}) - \mathbf{prox}_{\alpha h}(\mathbf{J}\mathbf{x}_t - \alpha \mathbf{J}\mathbf{y}_{t+1})\Big\|^2,
\end{aligned} \tag{6.75}$$

where the first line uses (6.73), the second line follows from Lemma 6.6.3(a), the third line is due to (6.74), and the last line uses Lemma 6.6.3(b). To proceed from (6.75), we observe that $\forall t \geq 1$,

$$\begin{aligned}
\Big\|\mathbf{prox}_{\alpha h}(\mathbf{x}_t - \alpha \mathbf{y}_{t+1}) - \mathbf{prox}_{\alpha h}(\mathbf{J}\mathbf{x}_t - \alpha \mathbf{J}\mathbf{y}_{t+1})\Big\|^2 &= \sum_{i=1}^n \Big\|\mathrm{prox}_{\alpha h}(\mathbf{x}_t^i - \alpha \mathbf{y}_{t+1}^i) - \mathrm{prox}_{\alpha h}(\overline{\mathbf{x}}_t - \alpha \overline{\mathbf{y}}_{t+1})\Big\|^2 \\
&\leq \sum_{i=1}^n \big\|\mathbf{x}_t^i - \overline{\mathbf{x}}_t - \alpha(\mathbf{y}_{t+1}^i - \overline{\mathbf{y}}_{t+1})\big\|^2 \\
&= \big\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t - \alpha(\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1})\big\|^2,
\end{aligned} \tag{6.76}$$

where the first and the second line uses (6.72) and Lemma 6.6.1 respectively. We then plug (6.76) into (6.75) to obtain: $\forall t \geq 1$ and $\forall \eta > 0$,

$$\begin{aligned}
\big\|\mathbf{x}_{t+1} - \mathbf{J}\mathbf{x}_{t+1}\big\|^2 &\leq \lambda^2 \big\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t - \alpha(\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1})\big\|^2 \\
&= \lambda^2 \big\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\big\|^2 + \lambda^2\alpha^2 \big\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\big\|^2 - 2\lambda^2 \big\langle \mathbf{x}_t - \mathbf{J}\mathbf{x}_t, \alpha(\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1})\big\rangle \\
&\leq \lambda^2 \big\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\big\|^2 + \lambda^2\alpha^2 \big\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\big\|^2 + 2\lambda^2 \big\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\big\| \big\|\alpha(\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1})\big\| \\
&\leq \lambda^2 (1+\eta) \big\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\big\|^2 + \lambda^2\alpha^2 (1+\eta^{-1}) \big\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\big\|^2,
\end{aligned} \tag{6.77}$$

where the third and the last line use the Cauchy-Schwarz and Young's inequality with parameter η respectively. Finally, setting $\eta = \frac{1-\lambda^2}{2\lambda^2}$ in (6.77) yields: $\forall t \geq 1$,

$$\big\|\mathbf{x}_{t+1} - \mathbf{J}\mathbf{x}_{t+1}\big\|^2 \leq \frac{1+\lambda^2}{2} \big\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\big\|^2 + \frac{\lambda^2\alpha^2(1+\lambda^2)}{1-\lambda^2} \big\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\big\|^2. \tag{6.78}$$

Applying Lemma 6.6.2 to (6.78), we have: $\forall T \geq 2$,

$$\sum_{t=1}^{T} \left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2 \leq \frac{2\lambda^2\alpha^2(1+\lambda^2)}{(1-\lambda^2)^2} \sum_{t=1}^{T-1} \left\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\right\|^2,$$

which finishes the proof of Lemma 6.6.6.

6.7.3 Proof of Lemma 6.6.7

6.7.3.1 Proof of Lemma 6.6.7(a)

We first recall that the gradient estimator $\mathbf{v}_t^i$ in Algorithm 8 takes the following form: $\forall t \geq 1$ and $i \in \mathcal{V}$,

$$\mathbf{v}_t^i := \frac{1}{b} \sum_{s=1}^{b} \nabla G_i\big(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t\big).$$

Observe that $\forall t \geq 1$,

$$\begin{aligned}
\mathbb{E}\Big[\left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2 \big| \mathcal{F}_t\Big] &= \mathbb{E}\left[\left\|\frac{1}{nb}\sum_{i=1}^{n}\sum_{s=1}^{b}\Big(\nabla G_i\big(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t\big) - \nabla f_i(\mathbf{x}_t^i)\Big)\right\|^2 \Bigg| \mathcal{F}_t\right] \\
&= \frac{1}{(nb)^2}\sum_{i=1}^{n}\sum_{s=1}^{b}\mathbb{E}\left[\left\|\nabla G_i\big(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t\big) - \nabla f_i(\mathbf{x}_t^i)\right\|^2 \big| \mathcal{F}_t\right] \\
&\leq \frac{1}{(nb)^2}\sum_{i=1}^{n}\sum_{s=1}^{b}\nu_i^2 \\
&= \frac{\nu^2}{nb},
\end{aligned}$$

where the second line uses Assumption 6.2.3 and the fact that $\mathbf{x}_t$ is $\mathcal{F}_t$-measurable and $\{\boldsymbol{\xi}_{i,s}^t : i \in \mathcal{V}, s \in [b]\}$ is independent of $\mathcal{F}_t$, while the third line is due to Assumption 6.2.4.

6.7.3.2 Proof of Lemma 6.6.7(b) and Lemma 6.6.7(c)

To facilitate the analysis, we first note that the gradient estimator $\mathbf{v}_t^i$ in both Algorithm 9 and 10 take the following form: $\forall i \in \mathcal{V}$ and $\forall t \geq 1$ such that $\mathrm{mod}(t, q) \neq 1$,

$$\mathbf{v}_t^i := \frac{1}{b}\sum_{s=1}^{b}\Big(\nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t) - \nabla G_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_{i,s}^t)\Big) + \mathbf{v}_{t-1}^i.$$

To simplify notation, we denote in this section that

$$\delta_t := \left\|\overline{\mathbf{v}}_t - \overline{\nabla \mathbf{f}}(\mathbf{x}_t)\right\|^2, \qquad \forall t \geq 1.$$

We establish an upper bound on δ_t that is applicable to both Algorithm 9 and 10.

Lemma 6.7.1. *Let Assumption 6.2.5 hold. Suppose that $T = Rq$ for some $R \in \mathbb{Z}^+$. Consider the iterates generated by Algorithm 9 or 10. Then we have: $\forall T \geq q$,*

$$\sum_{t=1}^{T}\mathbb{E}\big[\delta_t\big] \leq \frac{6L^2q}{n^2b}\sum_{t=1}^{T}\mathbb{E}\Big[\left\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\right\|^2\Big] + \frac{3qL^2\alpha^2}{nb}\sum_{t=1}^{T-1}\mathbb{E}\Big[\left\|\overline{\mathbf{g}}_t\right\|^2\Big] + q\sum_{z=1}^{R}\mathbb{E}\big[\delta_{(z-1)q+1}\big].$$

Proof. Consider any $t \geq 1$ such that $\mathrm{mod}(t, q) \neq 1$. For convenience, we define: $\forall i \in \mathcal{V}$,

$$\mathbf{d}_t^{i,s} := \nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t) - \nabla G_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_{i,s}^t), \qquad \mathbf{d}_t^i := \frac{1}{b}\sum_{s=1}^b \mathbf{d}_t^{i,s},$$

and we clearly have

$$\mathbb{E}\big[\mathbf{d}_t^{i,s}|\mathcal{F}_t\big] = \mathbb{E}\big[\mathbf{d}_t^i|\mathcal{F}_t\big] = \nabla f_i(\mathbf{x}_t^i) - \nabla f_i(\mathbf{x}_{t-1}^i). \tag{6.79}$$

As a consequence of (6.79) and of the independence between $\boldsymbol{\xi}_{i,s}^t$ and $\mathcal{F}_t$ for all $i \in \mathcal{V}$ and $s \in [b]$, we have

$$\mathbb{E}\Big[\Big\langle \mathbf{d}_t^i - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i), \mathbf{d}_t^r - \nabla f_r(\mathbf{x}_t^r) + \nabla f_r(\mathbf{x}_{t-1}^r)\Big\rangle|\mathcal{F}_t\Big] = 0, \tag{6.80}$$

whenever $i \neq r$, and

$$\mathbb{E}\Big[\Big\langle \mathbf{d}_t^{i,s} - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i), \mathbf{d}_t^{i,a} - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^r)\Big\rangle|\mathcal{F}_t\Big] = 0, \tag{6.81}$$

whenever $s \neq a$. Moreover, using the conditional variance decomposition with (6.79) gives: $\forall i \in \mathcal{V}$ and $s \in [b]$,

$$\mathbb{E}\Big[\big\|\mathbf{d}_t^{i,s} - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i)\big\|^2|\mathcal{F}_t\Big] \leq \mathbb{E}\Big[\big\|\mathbf{d}_t^{i,s}\big\|^2|\mathcal{F}_t\Big]. \tag{6.82}$$

By the update of $\mathbf{v}_t^i$, we observe that

$$\begin{aligned}
\mathbb{E}\big[\delta_t|\mathcal{F}_t\big] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{d}_t^i + \mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_t^i)\Big)\right\|^2 \Big|\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{d}_t^i - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) + \mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\Big)\right\|^2 \Big|\mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \Big(\mathbf{d}_t^i - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i)\Big)\right\|^2 \Big|\mathcal{F}_t\right] + \delta_{t-1} \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{d}_t^i - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2 \Big|\mathcal{F}_t\right] + \delta_{t-1} \\
&= \frac{1}{n^2 b^2}\sum_{i=1}^n\sum_{s=1}^b \mathbb{E}\left[\left\|\mathbf{d}_t^{i,s} - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i)\right\|^2 \Big|\mathcal{F}_t\right] + \delta_{t-1} \\
&\leq \frac{1}{n^2 b^2}\sum_{i=1}^n\sum_{s=1}^b \mathbb{E}\Big[\big\|\mathbf{d}_t^{i,s}\big\|^2|\mathcal{F}_t\Big] + \delta_{t-1},
\end{aligned} \tag{6.83}$$

where the third line uses (6.79), the fourth line uses (6.80), the fifth line uses (6.81), and the last line uses (6.82). We note that the mean-squared smoothness of $\nabla G(\cdot)$ implies that for all $i \in \mathcal{V}$ and $s \in [b]$,

$$\mathbb{E}\Big[\big\|\mathbf{d}_t^{i,s}\big\|^2\Big] \leq L^2\mathbb{E}\Big[\big\|\mathbf{x}_t^i - \mathbf{x}_{t-1}^i\big\|^2\Big]. \tag{6.84}$$

Applying (6.84) to (6.83) gives: for all $t \geq 1$ such that $\mathrm{mod}(t, q) \neq 1$,

$$\mathbb{E}\big[\delta_t\big] \leq \frac{L^2}{n^2 b}\mathbb{E}\Big[\big\|\mathbf{x}_t - \mathbf{x}_{t-1}\big\|^2\Big] + \mathbb{E}\big[\delta_{t-1}\big]. \tag{6.85}$$

For convenience, we define

$$\varphi_t := \left\lfloor \frac{t-1}{q} \right\rfloor, \qquad \forall t \geq 1.$$

It can be verified that

$$\varphi_t q + 1 \leq t \leq (\varphi_t + 1)q, \qquad \forall t \geq 1.$$

With the help of the above notations, we recursively apply (6.85) from t to $(\varphi_t q + 2)$ to obtain: for all $t \geq 1$ such that $\mathrm{mod}(t, q) \neq 1$,

$$\mathbb{E}\big[\delta_t\big] \leq \frac{L^2}{n^2 b}\sum_{j=\varphi_t q+2}^{t}\mathbb{E}\Big[\big\|\mathbf{x}_j - \mathbf{x}_{j-1}\big\|^2\Big] + \mathbb{E}\big[\delta_{\varphi_t q+1}\big]. \tag{6.86}$$

Summing up (6.86), we observe that $\forall z \geq 1$,

$$\begin{aligned}
\sum_{t=(z-1)q+1}^{zq}\mathbb{E}\big[\delta_t\big] &\leq \sum_{t=(z-1)q+2}^{zq}\left(\frac{L^2}{n^2 b}\sum_{j=\varphi_t q+2}^{t}\mathbb{E}\Big[\big\|\mathbf{x}_j - \mathbf{x}_{j-1}\big\|^2\Big] + \mathbb{E}\big[\delta_{\varphi_t q+1}\big]\right) + \mathbb{E}\big[\delta_{(z-1)q+1}\big] \\
&= \frac{L^2}{n^2 b}\sum_{t=(z-1)q+2}^{zq}\sum_{j=(z-1)q+2}^{t}\mathbb{E}\Big[\big\|\mathbf{x}_j - \mathbf{x}_{j-1}\big\|^2\Big] + q\mathbb{E}\big[\delta_{(z-1)q+1}\big] \\
&\leq \frac{L^2}{n^2 b}\sum_{t=(z-1)q+2}^{zq}\sum_{j=(z-1)q+2}^{zq}\mathbb{E}\Big[\big\|\mathbf{x}_j - \mathbf{x}_{j-1}\big\|^2\Big] + q\mathbb{E}\big[\delta_{(z-1)q+1}\big] \\
&= \frac{L^2(q-1)}{n^2 b}\sum_{j=(z-1)q+2}^{zq}\mathbb{E}\Big[\big\|\mathbf{x}_j - \mathbf{x}_{j-1}\big\|^2\Big] + q\mathbb{E}\big[\delta_{(z-1)q+1}\big],
\end{aligned} \tag{6.87}$$

where the second line uses the fact that $\varphi_t = z - 1$ when $(z-1)q + 1 \leq t \leq zq$ for all $z \geq 1$. Finally, we sum up (6.87) over z from 1 to R, we obtain: $\forall R \geq 1$,

$$\sum_{z=1}^{R}\sum_{t=(z-1)q+1}^{zq}\mathbb{E}\big[\delta_t\big] \leq \frac{L^2(q-1)}{n^2 b}\sum_{z=1}^{R}\sum_{j=(z-1)q+2}^{zq}\mathbb{E}\Big[\big\|\mathbf{x}_j - \mathbf{x}_{j-1}\big\|^2\Big] + q\sum_{z=1}^{R}\mathbb{E}\big[\delta_{(z-1)q+1}\big]. \tag{6.88}$$

Recall that $T = Eq$ and from (6.88) we obtain that $\forall T \geq q$,

$$\sum_{t=1}^{T}\mathbb{E}\big[\delta_t\big] \leq \frac{L^2(q-1)}{n^2 b}\sum_{t=2}^{T}\mathbb{E}\Big[\big\|\mathbf{x}_t - \mathbf{x}_{t-1}\big\|^2\Big] + q\sum_{z=1}^{R}\mathbb{E}\big[\delta_{(z-1)q+1}\big]. \tag{6.89}$$

Finally, we apply Lemma 6.6.4 to (6.89) to obtain: $\forall T \geq q$,

$$\sum_{t=1}^{T}\mathbb{E}\big[\delta_t\big] \leq \frac{6L^2(q-1)}{n^2 b}\sum_{t=1}^{T}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\big] + \frac{3L^2(q-1)\alpha^2}{nb}\sum_{t=1}^{T-1}\mathbb{E}\Big[\big\|\overline{\mathbf{g}}_t\big\|^2\Big] + q\sum_{z=1}^{R}\mathbb{E}\big[\delta_{(z-1)q+1}\big],$$

which finishes the proof. □

We observe that Lemma 6.6.7(b) follows from Lemma 6.7.1 by $\delta_{(z-1)q+1} = 0$ for all $z \geq 1$, while Lemma 6.6.7(c) follows by applying Lemma 6.6.7(a) to Lemma 6.7.1, i.e., $\mathbb{E}[\delta_{(z-1)q+1}] \leq \frac{\nu^2}{nB}$ for all $z \geq 1$.

6.7.4 Proof of Lemma 6.6.8

We first present a simple result that is useful for our later development.

Proposition 6.7.1. *Consider the iterates generated by Algorithm 7. The following inequality holds:* $\forall t \geq 1$,

$$\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 \leq \lambda^2\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 + 2\big\langle \mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\mathbf{v}_t - \mathbf{v}_{t-1}\big)\big\rangle. \tag{6.90}$$

Proof. Using the $\mathbf{y}$-update in Algorithm 7 and Lemma 6.6.3(a), we have: $\forall t \geq 1$,

$$\begin{aligned}
&\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 \\
&= \big\|\mathbf{W}^K\big(\mathbf{y}_t + \mathbf{v}_t - \mathbf{v}_{t-1}\big) - \mathbf{J}\mathbf{W}^K\big(\mathbf{y}_t + \mathbf{v}_t - \mathbf{v}_{t-1}\big)\big\|^2 \\
&= \big\|\mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t + \big(\mathbf{W}^K - \mathbf{J}\big)\big(\mathbf{v}_t - \mathbf{v}_{t-1}\big)\big\|^2 \\
&= \big\|\mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\big\|^2 + \big\|\big(\mathbf{W}^K - \mathbf{J}\big)\big(\mathbf{v}_t - \mathbf{v}_{t-1}\big)\big\|^2 + 2\big\langle \mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\mathbf{v}_t - \mathbf{v}_{t-1}\big)\big\rangle,
\end{aligned}$$

and the proof follows by using Lemma 6.6.3(c). □

6.7.4.1 Proof of Lemma 6.6.8(a)

Step 1: Decomposition. Recall that we are concerned with Algorithm 8 in this section. Conditioning (6.90) on $\mathcal{F}_t$, we have: $\forall t \geq 2$,

$$\begin{aligned}
&\mathbb{E}\big[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2|\mathcal{F}_t\big] \\
&\leq \lambda^2\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2|\mathcal{F}_t\big] + 2\big\langle \mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\nabla\mathbf{f}(\mathbf{x}_t) - \mathbf{v}_{t-1}\big)\big\rangle \\
&= \lambda^2\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2|\mathcal{F}_t\big] + 2\big\langle \mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\nabla\mathbf{f}(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\big)\big\rangle \\
&\quad + 2\big\langle \mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)\big\rangle \\
&= \lambda^2\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2|\mathcal{F}_t\big] + 2\big\langle \mathbf{W}^K\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\nabla\mathbf{f}(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\big)\big\rangle \\
&\quad + \underbrace{2\big\langle \mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)\big\rangle}_{=:A_t},
\end{aligned} \tag{6.91}$$

where the first line uses the fact that $\mathbf{x}_t$, $\mathbf{y}_t$, $\mathbf{v}_{t-1}$ are $\mathcal{F}_t$-measurable and also Assumption 6.2.3, while the last line uses Lemma 6.6.3(a). Towards the last term in (6.91), we observe that $\forall t \geq 2$ and $\forall \eta > 0$,

$$\begin{aligned}
A_t &\leq 2\big\|\mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\big\|\big\|\big(\mathbf{W}^K - \mathbf{J}\big)\big(\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)\big\| \\
&\leq 2\lambda\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|\lambda\|\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1})\| \\
&\leq 2\lambda\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|\lambda L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \\
&\leq \eta\lambda^2\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \eta^{-1}\lambda^2L^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2,
\end{aligned} \tag{6.92}$$

where the first line uses the Cauchy-Schwarz inequality, the second line uses Lemma 6.6.3, the third line uses the L-smoothness of each f_i, and last the line uses Young's inequality. Combining (6.92) and (6.91) leads to the following: $\forall t \geq 2$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2|\mathcal{F}_t\big] \leq (1+\eta)\lambda^2\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \eta^{-1}\lambda^2 L^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \\ + \lambda^2 \underbrace{\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2|\mathcal{F}_t\big]}_{=:B_t} + 2\underbrace{\big\langle \mathbf{W}^K\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\nabla\mathbf{f}(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\big)\big\rangle}_{=:C_t}. \end{aligned} \tag{6.93}$$

In the following, we bound B_t and C_t in (6.93) respectively.

Step 2: Controlling B_t. We decompose B_t as follows: $\forall t \geq 2$,

$$\begin{aligned} B_t &= \mathbb{E}\big[\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{x}_t) + \nabla\mathbf{f}(\mathbf{x}_t) - \mathbf{v}_{t-1}\|^2|\mathcal{F}_t\big] \\ &= \mathbb{E}\big[\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{x}_t)\|^2|\mathcal{F}_t\big] + \|\nabla\mathbf{f}(\mathbf{x}_t) - \mathbf{v}_{t-1}\|^2 \\ &\leq \mathbb{E}\big[\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{x}_t)\|^2|\mathcal{F}_t\big] + 2\|\nabla\mathbf{f}(\mathbf{x}_t) - \nabla\mathbf{f}(\mathbf{x}_{t-1})\|^2 + 2\|\nabla\mathbf{f}(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\|^2 \\ &\leq \mathbb{E}\big[\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{x}_t)\|^2|\mathcal{F}_t\big] + 2L^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 2\|\nabla\mathbf{f}(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\|^2, \end{aligned} \tag{6.94}$$

where the first line utilizes Assumption 6.2.3 and the fact that $\nabla\mathbf{f}(\mathbf{x}_t)$ and $\mathbf{v}_{t-1}$ are $\mathcal{F}_t$-measurable, while the last line uses the L-smoothness of each f_i. To proceed, we note that $\forall t \geq 1$,

$$\begin{aligned} \mathbb{E}\big[\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{x}_t)\|^2|\mathcal{F}_t\big] &= \sum_{i=1}^n \mathbb{E}\left[\left\|\frac{1}{b}\sum_{s=1}^b \nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t) - \nabla f_i(\mathbf{x}_t^i)\right\|^2 \Big|\mathcal{F}_t\right] \\ &= \frac{1}{b^2}\sum_{i=1}^n\sum_{s=1}^b \mathbb{E}\Big[\big\|\nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t) - \nabla f_i(\mathbf{x}_t^i)\big\|^2|\mathcal{F}_t\Big] \leq \frac{n\nu^2}{b}, \end{aligned} \tag{6.95}$$

where the second line uses the fact that $\mathbf{x}_t^i$ is $\mathcal{F}_t$-measurable and $\{\boldsymbol{\xi}_{i,1}^t, \cdots, \boldsymbol{\xi}_{i,b}^t, \mathcal{F}_t\}$ is an independent family for all $i \in \mathcal{V}$. Combining (6.94) and (6.95), we conclude that

$$\mathbb{E}\big[B_t\big] \leq 2L^2\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \frac{3n\nu^2}{b}, \qquad \forall t \geq 2. \tag{6.96}$$

Step 3: Controlling C_t. Towards C_t, we observe that $\forall t \geq 2$,

$$\begin{aligned} \mathbb{E}[C_t|\mathcal{F}_{t-1}] &= \mathbb{E}\Big[\Big\langle \mathbf{W}^{2K}\big(\mathbf{y}_{t-1} + \mathbf{v}_{t-1} - \mathbf{v}_{t-2}\big), \big(\mathbf{W}^K - \mathbf{J}\big)(\nabla\mathbf{f}(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1})\Big\rangle|\mathcal{F}_{t-1}\Big] \\ &= \mathbb{E}\Big[\Big\langle \mathbf{W}^{2K}\mathbf{v}_{t-1}, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\nabla\mathbf{f}(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\big)\Big\rangle|\mathcal{F}_{t-1}\Big] \\ &= \mathbb{E}\Big[\Big\langle \mathbf{W}^{2K}\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big), \big(\mathbf{J} - \mathbf{W}^K\big)\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)\Big\rangle|\mathcal{F}_{t-1}\Big], \end{aligned} \tag{6.97}$$

where the first line uses the $\mathbf{y}$-update in Algorithm 7, while the second and the last line use Assumption 6.2.3 with the $\mathcal{F}_{t-1}$-measurability of $\mathbf{y}_{t-1}$, $\mathbf{v}_{t-2}$ and $\nabla\mathbf{f}(\mathbf{x}_{t-1})$. To proceed, note that for all $t \geq 1$ we have

$$\mathbb{E}\Big[\Big\langle \mathbf{v}_t^i - \nabla f_i(\mathbf{x}_t^i), \mathbf{v}_t^r - \nabla f_i(\mathbf{x}_t^r)\Big\rangle|\mathcal{F}_t\Big] = 0, \tag{6.98}$$

whenever $i \neq r$. In light of (6.98), we proceed from (6.97) as follows: $\forall t \geq 2$,

$$\begin{aligned}
\mathbb{E}[C_t|\mathcal{F}_{t-1}] &= \mathbb{E}\Big[\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)^\top \big(\mathbf{J} - (\mathbf{W}^K)^\top \mathbf{W}^{2K}\big)\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)\big|\mathcal{F}_{t-1}\Big], \\
&= \mathbb{E}\Big[\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)^\top \mathrm{diag}\big(\mathbf{J} - (\mathbf{W}^\top)^K \mathbf{W}^{2K}\big)\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)\big|\mathcal{F}_{t-1}\Big] \\
&\leq \mathbb{E}\Big[\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)^\top \mathrm{diag}(\mathbf{J})\big(\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\big)\big|\mathcal{F}_{t-1}\Big] \\
&= \frac{1}{n}\mathbb{E}\big[\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{x}_{t-1})\|^2|\mathcal{F}_{t-1}\big] \\
&\leq \frac{\nu^2}{b},
\end{aligned} \tag{6.99}$$

where the first line uses Lemma 6.6.3(a), the second line uses (6.98), the third line uses the entry-wise nonnegativity of $\mathbf{W}$, and the last line uses (6.95). Therefore, we conclude from (6.99) that

$$\mathbb{E}[C_t] \leq \frac{\nu^2}{b}, \qquad \forall t \geq 2. \tag{6.100}$$

Step 4: Putting bounds together and refining. We substitute (6.96) and (6.100) into (6.93) to obtain: $\forall t \geq 2$,

$$\mathbb{E}\big[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2\big] \leq (1+\eta)\lambda^2\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] + (\eta^{-1}+2)\lambda^2L^2\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + (3\lambda^2 n + 2)\nu^2/b. \tag{6.101}$$

Setting $\eta = \frac{1-\lambda^2}{2\lambda^2}$, we have: $\forall t \geq 2$,

$$\mathbb{E}\big[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2\big] \leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] + \frac{2\lambda^2L^2}{1-\lambda^2}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \frac{(3\lambda^2 n + 2)\nu^2}{b}. \tag{6.102}$$

We then apply Lemma 6.6.2 to (6.102) to obtain: $\forall T \geq 2$,

$$\sum_{t=2}^{T+1}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] \leq \frac{2\mathbb{E}\big[\|\mathbf{y}_2 - \mathbf{J}\mathbf{y}_2\|^2\big]}{1-\lambda^2} + \frac{4\lambda^2L^2}{(1-\lambda^2)^2}\sum_{t=2}^{T}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \frac{2(T-1)(3\lambda^2 n + 2)\nu^2}{b(1-\lambda^2)}. \tag{6.103}$$

Since $\mathbf{y}_1 = \mathbf{v}_0 = \mathbf{0}_{np}$, we have

$$\begin{aligned}
\mathbb{E}\big[\|\mathbf{y}_2 - \mathbf{J}\mathbf{y}_2\|^2\big] = \mathbb{E}\big[\|(\mathbf{W}^K - \mathbf{J})\mathbf{v}_1\|^2\big] &\leq \lambda^2\mathbb{E}\big[\|\mathbf{v}_1\|^2\big] = \lambda^2\|\nabla\mathbf{f}(\mathbf{x}_1)\|^2 + \lambda^2\mathbb{E}\big[\|\mathbf{v}_1 - \nabla\mathbf{f}(\mathbf{x}_1)\|^2\big] \\
&\leq \lambda^2\|\nabla\mathbf{f}(\mathbf{x}_1)\|^2 + \lambda^2 n\nu^2/b,
\end{aligned} \tag{6.104}$$

where the second line uses Lemma 6.6.3(b), the third line uses Lemma 6.2.3, and the last line is due to (6.95). Finally, we apply (6.104) to (6.103) to obtain: $\forall T \geq 2$,

$$\sum_{t=2}^{T+1}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] \leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{4\lambda^2L^2}{(1-\lambda^2)^2}\sum_{t=2}^{T}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \frac{2T(3\lambda^2 n + 2)\nu^2}{b(1-\lambda^2)} + \frac{2\lambda^2 n\nu^2}{b(1-\lambda^2)}.$$

The proof of Lemma 6.6.8(a) follows by applying Lemma 6.6.4 to the above inequality with minor manipulations.

6.7.4.2 Proof of Lemma 6.6.8(b) and 6.6.8(c)

We first establish a gradient tracking error bound that is applicable to both Algorithm 9 and 10. For ease of exposition, we denote

$$\Upsilon_t := \|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{x}_t)\|^2, \qquad \forall t \geq 1. \tag{6.105}$$

Lemma 6.7.2. *Let Assumption 6.2.5 hold. Suppose that $T = Rq$ for some $R \in \mathbb{Z}^+$. Consider the iterates generated by Algorithm 9 or 10. Then we have: $\forall T \geq 2q$,*

$$\sum_{t=2}^{T+1} \mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] \leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{96\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=1}^{T} \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{J}\mathbf{x}_t\|^2\big] + \frac{48\lambda^2 n\alpha^2 L^2}{(1-\lambda^2)^2} \sum_{t=1}^{T-1} \mathbb{E}\big[\|\overline{\mathbf{g}}_t\|^2\big] + \frac{14\lambda^2}{(1-\lambda^2)^2} \sum_{z=0}^{R-1} \mathbb{E}\big[\Upsilon_{zq+1}\big]. \tag{6.106}$$

Proof. We first recall from (6.90) that $\forall t \geq 1$,

$$\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 \leq \lambda^2 \|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2 \|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 + 2\big\langle \mathbf{W}^K \mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\mathbf{v}_t - \mathbf{v}_{t-1}\big)\big\rangle. \tag{6.107}$$

In the first two steps, we refine (6.107) for $\mathrm{mod}(t,q) \neq 1$ and $\mathrm{mod}(t,q) = 1$ respectively.

Step 1: consider any $t \geq 2$ such that $\mathrm{mod}(t,q) \neq 1$. From (6.107), we observe that for all $\eta > 0$,

$$\begin{aligned}
\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 &\leq \lambda^2 \|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2 \|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 + 2\big\|\mathbf{W}^K \mathbf{y}_t - \mathbf{J}\mathbf{y}_t\big\|\big\|\big(\mathbf{W}^K - \mathbf{J}\big)\big(\mathbf{v}_t - \mathbf{v}_{t-1}\big)\big\| \\
&\leq \lambda^2 \|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2 \|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 + 2\lambda^2 \big\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\big\|\big\|\mathbf{v}_t - \mathbf{v}_{t-1}\big\| \\
&\leq \lambda^2(1+\eta)\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2(1+\eta^{-1})\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2,
\end{aligned} \tag{6.108}$$

where the first line uses Cauchy-Schwarz inequality, the second line uses Lemma 6.6.3, and the last uses Young's inequality. Setting $\eta = \frac{1-\lambda^2}{2\lambda^2}$ in (6.108) gives:

$$\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2 \leq \frac{1+\lambda^2}{2}\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \frac{\lambda^2(1+\lambda^2)}{1-\lambda^2}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 \tag{6.109}$$

Note that

$$\begin{aligned}
\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2\big] &= \sum_{i=1}^{n} \mathbb{E}\left[\left\|\frac{1}{b}\sum_{s=1}^{b}\Big(\nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t) - \nabla G_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_{i,s}^t)\Big)\right\|^2\right] \\
&\leq \frac{1}{b}\sum_{i=1}^{n}\sum_{s=1}^{b} \mathbb{E}\Big[\big\|\nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t) - \nabla G_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_{i,s}^t)\big\|^2\Big] \\
&\leq L^2 \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big],
\end{aligned} \tag{6.110}$$

where the last line uses the mean-squared smoothness. Applying (6.110) to (6.109), we obtain:

$$\mathbb{E}\big[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2\big] \leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] + \frac{2\lambda^2 L^2}{1-\lambda^2}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] \tag{6.111}$$

Step 2: consider any $t \geq 2$ such that $\mathrm{mod}(t,q) = 1$. In this case, we have $\mathbb{E}[\mathbf{v}_t|\mathcal{F}_t] = \nabla \mathbf{f}(\mathbf{x}_t)$. Taking the conditional expectation of (6.107) with respect to the filtration $\mathcal{F}_t$, we obtain

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2|\mathcal{F}_t\big] &\leq \lambda^2\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2|\mathcal{F}_t\big] \\ &\quad + 2\big\langle \mathbf{W}^K\mathbf{y}_t - \mathbf{J}\mathbf{y}_t, \big(\mathbf{W}^K - \mathbf{J}\big)\big(\nabla \mathbf{f}(\mathbf{x}_t) - \mathbf{v}_{t-1}\big)\big\rangle \\ &\leq \lambda^2(1+\eta)\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2 + \lambda^2\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2|\mathcal{F}_t\big] + \lambda^2\eta^{-1}\|\nabla \mathbf{f}(\mathbf{x}_t) - \mathbf{v}_{t-1}\|^2,\end{aligned} \tag{6.112}$$

where the first line uses the fact that $\mathbf{x}_t$ and $\mathbf{y}_t$ are $\mathcal{F}_t$-measurable and the second line follows a similar line of arguments as in (6.108). Setting $\eta = \frac{1-\lambda^2}{2\lambda^2}$ in (6.112), we obtain:

$$\mathbb{E}\big[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2\big] \leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] + \lambda^2\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2\big] + \frac{2\lambda^4}{1-\lambda^2}\mathbb{E}\big[\|\nabla \mathbf{f}(\mathbf{x}_t) - \mathbf{v}_{t-1}\|^2\big]. \tag{6.113}$$

We recall the definition of Υ_t in (6.105) and observe that

$$\begin{aligned}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 &= \|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{x}_t) + \nabla \mathbf{f}(\mathbf{x}_t) - \nabla \mathbf{f}(\mathbf{x}_{t-1}) + \nabla \mathbf{f}(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\|^2 \\ &\leq 3\Upsilon_t + 3\|\nabla \mathbf{f}(\mathbf{x}_t) - \nabla \mathbf{f}(\mathbf{x}_{t-1})\|^2 + 3\Upsilon_{t-1} \\ &\leq 3\Upsilon_t + 3L^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 3\Upsilon_{t-1},\end{aligned} \tag{6.114}$$

where the last uses the L-smoothness of each f_i. Similarly, we have

$$\begin{aligned}\|\nabla \mathbf{f}(\mathbf{x}_t) - \mathbf{v}_{t-1}\|^2 &= \|\nabla \mathbf{f}(\mathbf{x}_t) - \nabla \mathbf{f}(\mathbf{x}_{t-1}) + \nabla \mathbf{f}(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\|^2 \\ &\leq 2L^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + 2\Upsilon_{t-1}.\end{aligned} \tag{6.115}$$

Plugging (6.114) and (6.115) into (6.113) gives

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2\big] &\leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] + \left(3\lambda^2 + \frac{4\lambda^4}{1-\lambda^2}\right)L^2\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] \\ &\quad + 3\lambda^2\mathbb{E}\big[\Upsilon_t\big] + \left(3\lambda^2 + \frac{4\lambda^4}{1-\lambda^2}\right)\mathbb{E}\big[\Upsilon_{t-1}\big] \\ &\leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] + \frac{4\lambda^2L^2}{1-\lambda^2}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + 3\lambda^2\mathbb{E}\big[\Upsilon_t\big] + \frac{4\lambda^2\mathbb{E}\big[\Upsilon_{t-1}\big]}{1-\lambda^2}.\end{aligned} \tag{6.116}$$

Step 3: combining step 1 and step 2. Combining (6.111) and (6.116), we obtain: $\forall t \geq 2$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{y}_{t+1} - \mathbf{J}\mathbf{y}_{t+1}\|^2\big] &\leq \frac{1+\lambda^2}{2}\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] + \frac{4\lambda^2L^2}{1-\lambda^2}\mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] \\ &\quad + \mathbb{1}_{\{\mathrm{mod}\,(t,q)=1\}}\left(3\lambda^2\mathbb{E}\big[\Upsilon_t\big] + \frac{4\lambda^2\mathbb{E}[\Upsilon_{t-1}]}{1-\lambda^2}\right).\end{aligned} \tag{6.117}$$

Let $T = Rq$ for some $R \in \mathbb{Z}^+$. We apply Lemma 6.6.2 to (6.117) to obtain: $\forall T \geq 2q$,

$$\begin{aligned}
&\sum_{t=2}^{T+1} \mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] \\
&\leq \frac{2\mathbb{E}\big[\|\mathbf{y}_2 - \mathbf{J}\mathbf{y}_2\|^2\big]}{1-\lambda^2} + \frac{8\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=2}^{T} \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \sum_{t=2}^{T} \mathbb{1}_{\{\mathrm{mod}\ (t,q)=1\}} \left(\frac{6\lambda^2 \mathbb{E}[\Upsilon_t]}{1-\lambda^2} + \frac{8\lambda^2 \mathbb{E}[\Upsilon_{t-1}]}{(1-\lambda^2)^2} \right) \\
&= \frac{2\mathbb{E}\big[\|\mathbf{y}_2 - \mathbf{J}\mathbf{y}_2\|^2\big]}{1-\lambda^2} + \frac{8\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=2}^{T} \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \sum_{z=1}^{R-1} \left(\frac{6\lambda^2}{1-\lambda^2} \mathbb{E}\big[\Upsilon_{zq+1}\big] + \frac{8\lambda^2}{(1-\lambda^2)^2} \mathbb{E}\big[\Upsilon_{zq}\big] \right).
\end{aligned} \tag{6.118}$$

Note that

$$\mathbb{E}\big[\|\mathbf{y}_2 - \mathbf{J}\mathbf{y}_2\|^2\big] = \mathbb{E}\big[\|(\mathbf{W}^K - \mathbf{J})\mathbf{v}_1\|^2\big] \leq \lambda^2 \mathbb{E}\big[\|\mathbf{v}_1\|^2\big] = \lambda^2 \|\nabla \mathbf{f}(\mathbf{x}_1)\|^2 + \lambda^2 \mathbb{E}\big[\Upsilon_1\big] \tag{6.119}$$

Applying (6.119) to (6.118) gives the following: $\forall T \geq 2q$,

$$\begin{aligned}
&\sum_{t=2}^{T+1} \mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] \\
&\leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{8\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=2}^{T} \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \frac{2\lambda^2}{1-\lambda^2} \mathbb{E}[\Upsilon_1] + \frac{6\lambda^2}{1-\lambda^2} \sum_{z=1}^{R-1} \mathbb{E}\big[\Upsilon_{zq+1}\big] + \frac{8\lambda^2}{(1-\lambda^2)^2} \sum_{z=1}^{R-1} \mathbb{E}\big[\Upsilon_{zq}\big] \\
&\leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{8\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=2}^{T} \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \frac{6\lambda^2}{1-\lambda^2} \sum_{z=0}^{R-1} \mathbb{E}\big[\Upsilon_{zq+1}\big] + \frac{8\lambda^2}{(1-\lambda^2)^2} \sum_{z=1}^{R-1} \mathbb{E}\big[\Upsilon_{zq}\big]
\end{aligned} \tag{6.120}$$

Step 4: bounding $\mathbb{E}\big[\Upsilon_t\big]$. The derivations in this step essentially repeat the proof of Lemma 6.7.1. Recall the definition of Υ_t in (6.105). Consider any $t \geq 1$ such that $\mathrm{mod}(t,q) \neq 1$ and define: $\forall i \in \mathcal{V}$,

$$\mathbf{d}_t^{i,s} := \nabla G_i(\mathbf{x}_t^i, \boldsymbol{\xi}_{i,s}^t) - \nabla G_i(\mathbf{x}_{t-1}^i, \boldsymbol{\xi}_{i,s}^t), \qquad \mathbf{d}_t^i := \frac{1}{b} \sum_{s=1}^{b} \mathbf{d}_t^{i,s}.$$

By the update of $\mathbf{v}_t^i$, we observe that

$$\begin{aligned}
\mathbb{E}\big[\Upsilon_t | \mathcal{F}_t\big] = \sum_{i=1}^{n} \mathbb{E}\Big[\big\|\mathbf{d}_t^i + \mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_t^i)\big\|^2 \,\big|\mathcal{F}_t\Big] &= \sum_{i=1}^{n} \mathbb{E}\Big[\big\|\mathbf{d}_t^i - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i) + \mathbf{v}_{t-1}^i - \nabla f_i(\mathbf{x}_{t-1}^i)\big\|^2 \,\big|\mathcal{F}_t\Big] \\
&= \sum_{i=1}^{n} \mathbb{E}\Big[\big\|\mathbf{d}_t^i - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i)\big\|^2 \,\big|\mathcal{F}_t\Big] + \Upsilon_{t-1} \\
&= \frac{1}{b^2} \sum_{i=1}^{n} \sum_{s=1}^{b} \mathbb{E}\Big[\big\|\mathbf{d}_t^{i,s} - \nabla f_i(\mathbf{x}_t^i) + \nabla f_i(\mathbf{x}_{t-1}^i)\big\|^2 \big|\mathcal{F}_t\Big] + \Upsilon_{t-1} \\
&\leq \frac{1}{b^2} \sum_{i=1}^{n} \sum_{s=1}^{b} \mathbb{E}\Big[\big\|\mathbf{d}_t^{i,s}\big\|^2 \big|\mathcal{F}_t\Big] + \Upsilon_{t-1},
\end{aligned} \tag{6.121}$$

where the above derivations follow a very similar line of arguments as in (6.83) and thus we omit the detailed explanations. Taking the expectation of (6.121) and using the mean-squared smoothness of $\nabla G(\cdot, \boldsymbol{\xi})$, we obtain

$$\mathbb{E}\big[\Upsilon_t\big] \leq \frac{L^2}{b} \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \mathbb{E}\big[\Upsilon_{t-1}\big]. \tag{6.122}$$

For convenience, we define $\varphi_t := \left\lfloor \frac{t-1}{q} \right\rfloor, \forall t \geq 1$. It can be verified that $\varphi_t q + 1 \leq t \leq (\varphi_t + 1)q, \forall t \geq 1$. Recursively applying (6.122) from t to $(\varphi_t + 1)$, we obtain

$$\mathbb{E}\big[\Upsilon_t\big] \leq \frac{L^2}{b} \sum_{j=\varphi_t+1}^{t} \mathbb{E}\big[\|\mathbf{x}_j - \mathbf{x}_{j-1}\|^2\big] + \mathbb{E}\big[\Upsilon_{\varphi_t}\big]. \tag{6.123}$$

In particular, taking $t = zq$ for some $z \in \mathbb{Z}^+$ in (6.123) gives

$$\mathbb{E}\big[\Upsilon_{zq}\big] \leq \frac{L^2}{b} \sum_{j=(z-1)q+2}^{zq} \mathbb{E}\big[\|\mathbf{x}_j - \mathbf{x}_{j-1}\|^2\big] + \mathbb{E}\big[\Upsilon_{(z-1)q+1}\big]. \tag{6.124}$$

We sum up (6.124) over z from 1 to R to obtain:

$$\begin{aligned}\sum_{z=1}^{R} \mathbb{E}\big[\Upsilon_{zq}\big] &\leq \frac{L^2}{b} \sum_{z=1}^{R} \sum_{j=(z-1)q+2}^{zq} \mathbb{E}\big[\|\mathbf{x}_j - \mathbf{x}_{j-1}\|^2\big] + \sum_{z=1}^{R} \mathbb{E}\big[\Upsilon_{(z-1)q+1}\big] \\ &\leq \frac{L^2}{b} \sum_{t=2}^{T} \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \sum_{z=0}^{R-1} \mathbb{E}\big[\Upsilon_{zq+1}\big]. \end{aligned} \tag{6.125}$$

Step 5: putting bounds together. Applying (6.125) to (6.120) gives the following: $\forall T \geq 2q$,

$$\begin{aligned}&\sum_{t=2}^{T+1} \mathbb{E}\big[\|\mathbf{y}_t - \mathbf{J}\mathbf{y}_t\|^2\big] \\ &\leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \left(1 + \frac{1}{b}\right) \frac{8\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=2}^{T} \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \frac{6\lambda^2}{1-\lambda^2} \sum_{z=0}^{R-1} \mathbb{E}\big[\Upsilon_{zq+1}\big] + \frac{8\lambda^2}{(1-\lambda^2)^2} \sum_{z=0}^{R-1} \mathbb{E}\big[\Upsilon_{zq+1}\big] \\ &\leq \frac{2\lambda^2 n\zeta^2}{1-\lambda^2} + \frac{16\lambda^2 L^2}{(1-\lambda^2)^2} \sum_{t=2}^{T} \mathbb{E}\big[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2\big] + \frac{14\lambda^2}{(1-\lambda^2)^2} \sum_{z=0}^{R-1} \mathbb{E}\big[\Upsilon_{zq+1}\big]. \end{aligned} \tag{6.126}$$

Plugging Lemma 6.6.4 into (6.126) finishes the proof. □

Note that Lemma 6.6.8(b) follows from (6.106) by $\Upsilon_{zq+1} \leq n\nu^2/B$ for all $z \in \mathbb{Z}^+$, while Lemma 6.6.8(c) follows from (6.106) by $\Upsilon_{zq+1} = 0$ for all $z \in \mathbb{Z}^+$.

6.8 Conclusion

In this chapter, we have studied decentralized non-convex non-smooth composite problems under expected or empirical risk. This formulation generalizes the problems considered in the previous chapters by adding an extended valued, convex, possibly non-smooth regularization term to the risk functions. In this context, we propose the first provably efficient decentralized proximal gradient framework whose instances achieve gradient and communication complexities that match the centralized optimal methods for the corresponding problem classes. Several technical lemmas in the convergence analysis are of independent interest and helpful to analyze other decentralized algorithms based on similar principles.

Chapter 7

Epilogue

We now revisit the major contributions of this thesis. In this thesis, we study several fundamental classes of decentralized stochastic non-convex optimization and learning problems over heterogeneous data, with emphasis on machine learning and signal processing applications. In particular, we propose a family of provably fast and robust decentralized algorithms with the help of gradient tracking, variance reduction, mini-batch stochastic gradient, and multi-round accelerated consensus techniques. We further prove that these decentralized algorithms, with appropriate parameters, achieve optimal gradient and communication complexities for the corresponding problem classes. In light of these convergence results, we provide a theoretical justification that decentralized optimization methods can outperform the corresponding centralized optimal ones in regimes of practical significance, e.g., when the communication network is of low bandwidth or the power budget at each node is limited. Throughout the thesis, we also provide numerical illustrations to validate our theoretical results. The convergence analysis and several intermediate technical results developed in this thesis are of independent interest and may be helpful to address other decentralized non-convex formulations. In the following, we recap the contributions of each chapter.

- **Chapter 2: smooth strongly-convex empirical risk minimization**. In this chapter, we have proposed a novel framework for constructing variance-reduced decentralized stochastic first-order methods over undirected and weight-balanced directed graphs that hinge on gradient tracking techniques. In particular, we derive under this framework decentralized versions of the centralized `SAGA` and `SVRG` algorithms, namely **`GT-SAGA`** and **`GT-SVRG`**, that achieve accelerated linear convergence for smooth and strongly convex functions compared with existing decentralized stochastic first-order methods. We have further shown that in the big-data regimes, **`GT-SAGA`** and **`GT-SVRG`** achieve non-asymptotic, linear speedups in terms of the number of nodes compared with centralized `SAGA` and `SVRG`. Extensive numerical experiments based on real-world datasets are provided to validate our theoretical findings.

- **Chapter 3: smooth non-convex empirical risk minimization.** In this chapter, we have proposed two decentralized variance-reduced first-order gradient methods, `GT-SARAH` and `GT-SAGA`, to minimize a finite-sum of N smooth non-convex cost functions equally distributed over a decentralized network of n nodes. With appropriate algorithmic parameters, `GT-SARAH` achieves significantly improved gradient complexity compared with the existing decentralized stochastic gradient methods. In particular, in a big-data regime $n = \mathcal{O}(N^{1/2}(1-\lambda)^3)$, the gradient complexity of `GT-SARAH` reduces to $\mathcal{O}(N^{1/2}L\epsilon^{-2})$ which matches the centralized lower bound for this problem class, where L is the smoothness parameter and $(1-\lambda)$ is the spectral gap of the network weight matrix. Furthermore, `GT-SARAH` in this regime achieves non-asymptotic linear speedup compared with the centralized optimal approaches such as `SPIDER` [49,50] and `SARAH` [48] that perform all gradient computations on a single machine. Compared with the implementations of `SPIDER` and `SARAH` over server-worker architectures [132], the decentralized `GT-SARAH` enjoys the same non-asymptotic linear speedup in terms of the gradient complexity, however, admits sparser and more flexible communication topology and thus reduced total run time. In a large-scale network regime like the Internet of Things (IoT) where the number of nodes and the spectral gap of the network are considerably large, we show that `GT-SAGA` provably achieves faster convergence rate and more practical implementation than `GT-SARAH` and other existing decentralized algorithms.

- **Chapter 4: general smooth expected risk minimization.** In this chapter, we have comprehensively improved the existing convergence results of decentralized stochastic first-order methods based on gradient tracking for online stochastic non-convex problems. In particular, for both constant and decaying step-sizes, we systematically develop the conditions under which the performance of `GT-DSGD` matches that of the centralized minibatch `SGD` for both general non-convex functions and non-convex functions that further satisfy the PL condition. Specifically, we show that if the required error tolerance of the solution is small enough, then `GT-DSGD` matches the centralized lower bound for these problem classes. Our convergence results significantly improve upon the existing theory, which suggests that `GT-DSGD` is strictly worse than the centralized minibatch `SGD`. For a family of stochastic approximation step-sizes, we establish, for the first time, the optimal global sublinear convergence to an optimal solution on almost every sample path of `GT-DSGD`, when the global function satisfies the PL condition.

- **Chapter 5: non-convex expected risk minimization with mean-squared smoothness.** In this chapter, we have investigated decentralized stochastic optimization to minimize smooth non-convex cost functions distributed over networked nodes. Under the assumption that the stochastic gradient satisfies the mean-squared smoothness condition, we propose `GT-HSGD`, a novel single-loop decentralized algorithm that leverages local hybrid variance-reduced estimators and gradient tracking to achieve

provably fast convergence rate and robust performance under heterogenous data. Compared with the existing decentralized online variance-reduced methods, `GT-HSGD` achieves a lower gradient complexity with a more practical implementation. We further show that `GT-HSGD` matches the centralized lower bound for this problem class, when the required error tolerance is small enough, leading to a linear speedup with respect to the centralized optimal methods that are implemented on a single machine.

- **Chapter 6: non-convex non-smooth composite problems.** In this chapter, we have developed a unified stochastic proximal gradient tracking framework, called `ProxGT`, for decentralized *non-convex non-smooth composite* minimization problems with empirical or expected risk. Here, a decentralized network of nodes collaborates to find a stationary point of the average of smooth non-convex local costs plus an extended valued, convex, possibly non-smooth global regularizer that enfores additional structures to the problem. This composite formulation is considerably general and covers many practical applications of interest such as sparse and constrained optimization problems. We specifically develop instances of `ProxGT` that achieve optimal gradient and communication complexities simultaneously for different problem classes. In the convergence analysis, we establish a novel decentralized stochastic proximal descent inequality and a new proximal consensus error bound which may be of independent interest and can be helpful to analyze other decentralized stochastic algorithms based on similar principles such as proximal variants of DSGD.

Bibliography

[1] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011. xi, 26

[2] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 5330–5340. xi, 5, 8, 9, 15, 17, 18, 49, 50, 52, 56, 59, 82, 87, 107, 113, 116, 139, 140, 142, 148, 167

[3] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "D^2: Decentralized training over decentralized data," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 4848–4856. xi, 9, 11, 18, 50, 51, 52, 55, 56, 59, 107, 140, 142, 144, 148, 167

[4] R. Xin, U. A. Khan, and S. Kar, "An improved convergence analysis for decentralized online stochastic non-convex optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 1842–1858, 2021. xi, 4, 9, 11, 18, 31, 50, 51, 52, 59, 79, 81, 82, 85, 90, 96, 140, 142, 144, 148, 173

[5] B. T. Polyak, "Introduction to optimization," *Inc., Publications Division, New York*, vol. 1, 1987. 1, 81, 90, 109, 110, 121, 122, 128

[6] Y. Nesterov, *Lectures on convex optimization*, vol. 137, Springer, 2018. 1, 15, 63, 64, 155

[7] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018. 1, 3, 15, 49, 50, 55, 63, 64, 107, 109, 111, 115, 122, 135, 167

[8] G. Lan, *First-order and Stochastic Optimization Methods for Machine Learning*, Springer, 2020. 1, 171, 172, 173

[9] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, SIAM, 2001. 1

[10] A. Beck, *First-order methods in optimization*, SIAM, 2017. 1, 92, 93, 171, 173, 182, 190, 191

[11] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997. 1

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. 2

[13] T. Hastie, R. Tibshirani, J. H Friedman, and J. H Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2, Springer, 2009. 2, 15

[14] R. Xin, S. Pu, A. Nedić, and U. A. Khan, "A general framework for decentralized optimization with first-order methods," *P. IEEE*, vol. 108, no. 11, pp. 1869–1889, 2020. 3, 5, 10, 11, 107, 110, 140, 148, 168

[15] T. Chang, M. Hong, H. Wai, X. Zhang, and S. Lu, "Distributed learning in the nonconvex world: From batch data to streaming and beyond," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 26–38, 2020. 3, 5, 9, 11, 50, 59

[16] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020. 3, 5

[17] V. Smith, S. Forte, M. Chenxin, M. Takac, M. I. Jordan, and M. Jaggi, "CoCoA: A general framework for communication-efficient distributed optimization," *J. Mach. Learn. Res.*, vol. 18, pp. 230, 2018. 3

[18] S. Kar and J. M. F. Moura, "Consensus+ innovations distributed inference over networks: cooperation and sensing in networked systems," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 99–109, 2013. 3

[19] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks—Part I: transient analysis," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3487–3517, 2015. 3, 4, 6, 8, 17, 19, 49, 139, 140

[20] K. Yuan, S. A. Alghunaim, B. Ying, and A. H. Sayed, "On the influence of bias-correction on distributed stochastic optimization," *IEEE Trans. Signal Process.*, 2020. 4, 9, 11, 18, 19, 30, 50, 51, 107, 109, 140

[21] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3575–3605, 2012. 4, 6, 7, 10, 15, 17, 19, 139, 140, 166

[22] A. Mokhtari and A. Ribeiro, "DSA: Decentralized double stochastic averaging gradient algorithm," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2165–2199, 2016. 4, 18, 25, 51, 79, 85, 140

[23] K. Yuan, B. Ying, J. Liu, and A. H. Sayed, "Variance-reduced stochastic learning by networked agents under random reshuffling," *IEEE Trans. Signal Process.*, vol. 67, no. 2, pp. 351–366, 2018. 4, 18, 30, 51, 79, 85, 140

[24] R. Xin, U. A. Khan, and S. Kar, "Variance-reduced decentralized stochastic optimization with accelerated convergence," *IEEE Trans. Signal Process.*, vol. 68, pp. 6255–6271, 2020. 4, 51, 78, 79, 85, 86, 90, 120, 150

[25] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008. 5

[26] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011. 5

[27] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *P. IEEE*, vol. 106, no. 5, pp. 953–976, 2018. 5, 7, 11, 26, 49, 55, 59, 60, 86, 116, 148, 168, 172, 173, 179

[28] B. Ying, K. Yuan, Y. Chen, H. Hu, P. Pan, and W. Yin, "Exponential graph is provably efficient for decentralized deep training," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 13975–13987, 2021. 5, 15

[29] K. Yuan, Y. Chen, X. Huang, Y. Zhang, P. Pan, Y. Xu, and W. Yin, "DecentLaM: Decentralized momentum SGD for large-batch deep training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3029–3039. 5, 15, 167

[30] D. Alistarh, Z. Allen-Zhu, and J. Li, "Byzantine stochastic gradient descent," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018. 5

[31] R. Xin, S. Kar, and U. A. Khan, "Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 102–113, 2020. 5, 11, 17, 18, 30, 49, 51, 59, 140, 166

[32] X. Mao, K. Yuan, Y. Hu, Y. Gu, A. H. Sayed, and W. Yin, "Walkman: A communication-efficient random-walk algorithm for decentralized optimization," *IEEE Trans. Signal Process.*, vol. 68, pp. 2513–2528, 2020. 6

[33] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48, 2009. 6, 7, 10, 17, 18, 139, 166

[34] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, 1986. 6, 10, 17, 139

[35] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *P. IEEE*, vol. 95, no. 1, pp. 215–233, 2007. 6, 7, 10

[36] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, 2012. 7, 10, 23, 39, 46, 55, 61, 73, 100, 111, 126, 172

[37] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, 2012. 7, 8, 10, 49, 107, 110, 166

[38] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM J. Optim.*, vol. 26, no. 3, pp. 1835–1854, 2016. 8, 10, 49

[39] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, no. 3, pp. 516–545, 2010. 8, 17, 19, 49, 107

[40] J. Wang, V. Tantia, N. Ballas, and M. Rabbat, "SlowMo: Improving communication-efficient distributed SGD with slow momentum," *arXiv preprint arXiv:1910.00643*, 2019. 8, 49, 50, 167

[41] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 97: 344–353. 8, 9, 30, 49, 50, 86, 107, 139, 144, 148, 167

[42] B. Swenson, R. Murray, S. Kar, and H. V. Poor, "Distributed stochastic gradient descent and convergence to local minima," *J. Mach. Learn. Res.*, 2022. 8, 9, 11, 49, 50, 107, 168

[43] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—Part II: Polynomial escape from saddle-points," *IEEE Trans. Signal Process.*, vol. 69, pp. 1257–1270, 2021. 8, 9, 11, 18, 49, 50, 107, 140

[44] Y. Wang, W. Zhao, Y. Hong, and M. Zamani, "Distributed subgradient-free stochastic optimization algorithm for nonsmooth convex functions over time-varying networks," *SIAM J. Control Optim.*, vol. 57, no. 4, pp. 2821–2842, 2019. 8, 49

[45] S. Pu and A. Garcia, "Swarming for faster convergence in stochastic optimization," *SIAM J. Control Optim.*, vol. 56, no. 4, pp. 2997–3020, 2018. 8, 49

[46] D. Yuan, D. W. Ho, and Y. Hong, "On convergence rate of distributed stochastic gradient algorithm for convex optimization with inequality constraints," *SIAM J. Control Optim.*, vol. 54, no. 5, pp. 2872–2892, 2016. 8, 49

[47] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2341–2368, 2013. 9, 49, 50

[48] N. H. Pham, L. M. Nguyen, D. T. Phan, and Q. Tran-Dinh, "ProxSARAH: an efficient algorithmic framework for stochastic composite nonconvex optimization," *J. Mach. Learn. Res.*, vol. 21, no. 110, pp. 1–48, 2020. 9, 50, 51, 52, 53, 54, 56, 58, 63, 105, 140, 141, 147, 169, 175, 177, 178, 206

[49] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 689–699. 9, 50, 51, 52, 53, 54, 55, 56, 58, 63, 105, 140, 141, 144, 147, 169, 174, 175, 206

[50] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh, "Spiderboost and momentum: Faster variance reduction algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 2403–2413. 9, 50, 51, 52, 54, 56, 58, 63, 105, 140, 141, 144, 147, 169, 174, 175, 177, 178, 206

[51] M. Zhu and S. Martínez, "Discrete-time dynamic average consensus," *Automatica*, vol. 46, no. 2, pp. 322–329, 2010. 9, 10, 17, 20, 31, 110, 174

[52] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Syst. Lett.*, vol. 2, no. 3, pp. 315–320, 2018. 9, 10, 18, 30, 31, 107

[53] S. Pu, W. Shi, J. Xu, and A. Nedić, "A push-pull gradient method for distributed optimization in networks," in *Proc. IEEE Conf. Decis. Control*, Dec. 2018, pp. 3385–3390. 9, 10, 18, 30, 31, 107

[54] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2597–2633, 2017. 9, 10, 11, 17, 18, 20, 31, 51, 53, 63, 107, 114, 140

[55] P. Di Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Trans. Signal Inf. Process. Netw. Process.*, vol. 2, no. 2, pp. 120–136, 2016. 9, 10, 11, 17, 18, 20, 50, 51, 53, 78, 79, 90, 107, 140, 144, 150, 168, 173

[56] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control. Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, 2017. 9, 10, 11, 17, 18, 20, 30, 31, 51, 53, 55, 63, 79, 90, 107, 114, 120, 140, 150

[57] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1646–1654. 9, 17, 20, 21, 24, 25, 31, 51, 78, 79, 81, 90

[58] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takac, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 2613–2621. 9, 17, 20, 50, 51, 140, 144

[59] D. Liu, L. M. Nguyen, and Q. Tran-Dinh, "An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization," *arXiv preprint arXiv:2008.09055*, 2020. 9, 140, 141, 144, 147, 174

[60] S. J. Reddi, S. Sra, B. Póczos, and A. Smola, "Fast incremental method for smooth nonconvex optimization," in *Proc. IEEE Conf. Decis. Control*, 2016, pp. 1971–1977. 9, 78, 79, 81, 83, 85

[61] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 315–323. 9, 17, 20, 24, 31

[62] Z. Allen-Zhu and E. Hazan, "Variance reduction for faster non-convex optimization," in *Proc. 33th Int. Conf. Mach. Learn.*, 2016, pp. 699–707. 9, 51, 174

[63] S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *Proc. 33th Int. Conf. Mach. Learn.*, 2016, pp. 314–323. 9, 51

[64] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM J. Optim.*, vol. 24, no. 4, pp. 2057–2075, 2014. 9, 51

[65] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *Proc. IEEE Conf. Decis. Control*, 2015, pp. 2055–2060. 9, 10, 11, 18, 50, 53, 78, 79, 107, 114, 140, 144, 173

[66] S. A. Alghunaim, E. Ryu, K. Yuan, and A. H. Sayed, "Decentralized proximal gradient algorithms with linear convergence rates," *IEEE Trans. Autom. Control*, 2020. 10, 11, 140

[67] S. Pu and A. Nedich, "Distributed stochastic gradient tracking methods," *Math. Program.*, pp. 1–49, 2020. 10, 11, 18, 30, 31, 51, 79, 81, 96, 107, 108, 109, 110, 113, 114, 115, 120, 137, 140, 144, 148, 150, 152, 173

[68] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Optim.*, vol. 25, no. 2, pp. 944–966, 2015. 11, 18, 25, 51, 107, 140

[69] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4494–4506, 2019. 11, 51, 114, 140

[70] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, "DLM: Decentralized linearized alternating direction method of multipliers," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 4051–4064, 2015. 11, 18

[71] M. Maros and J. Jaldén, "A geometrically converging dual method for distributed optimization over time-varying graphs," *IEEE Trans. Autom. Control*, 2020. 11

[72] H. Wai, J. Lafond, A. Scaglione, and E. Moulines, "Decentralized frank–wolfe algorithm for convex and nonconvex problems," *IEEE Trans. Autom. Control*, vol. 62, no. 11, pp. 5522–5537, 2017. 11, 51, 168

[73] D. Jakovetić, "A unification and generalization of exact distributed first-order methods," *IEEE Trans. Signal Inf. Process. Netw. Process.*, vol. 5, no. 1, pp. 31–46, 2018. 11, 18, 53, 140

[74] J. Xu, Y. Tian, Y. Sun, and G. Scutari, "Distributed algorithms for composite optimization: unified framework and convergence analysis," *IEEE Trans. Signal Process.*, vol. 69, pp. 3555–3570, 2021. 11, 140, 168

[75] X. Wu and J. Lu, "A unifying approximate method of multipliers for distributed composite optimization," *IEEE Trans. Autom. Control*, 2022. 11, 168

[76] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić, "A dual approach for optimal algorithms in distributed optimization over networks," in *2020 Information Theory and Applications Workshop (ITA)*. IEEE, 2020, pp. 1–37. 11

[77] K. Seaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3027–3036. 11, 178

[78] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "A decentralized second-order method with exact linear convergence rate for consensus optimization," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 4, pp. 507–522, 2016. 11

[79] M. Eisen, A. Mokhtari, and A. Ribeiro, "Decentralized quasi-newton methods," *IEEE Trans. on Sig. Process.*, vol. 65, no. 10, pp. 2613–2628, May 2017. 11

[80] A. Mokhtari, Q. Ling, and A. Ribeiro, "Network Newton distributed optimization methods," *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 146–161, Jan 2017. 11

[81] H. Wai, N. M. Freris, A. Nedić, and A. Scaglione, "SUCAG: stochastic unbiased curvature-aided gradient method for distributed optimization," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 1751–1756. 11

[82] F. Mansoori and E. Wei, "Superlinearly convergent asynchronous distributed network newton method," in *Proc. IEEE Conf. Decis. Control*, 2017, pp. 2874–2879. 11

[83] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Puschel, "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2718–2723, May 2013. 11

[84] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, 2014. 11

[85] E. Wei and A. Ozdaglar, "On the $o(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers," in *IEEE Global Conference on Signal and Information Processing*. IEEE, 2013, pp. 551–554. 11

[86] M. Maros and J. Jaldén, "On the Q-linear convergence of distributed generalized ADMM under non-strongly convex function components," *IEEE Trans. Signal Inf. Process. Netw.*, 2019. 11

[87] H. Sun and M. Hong, "Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms," *IEEE Trans. Signal process.*, vol. 67, no. 22, pp. 5912–5928, 2019. 11, 55

[88] M. Hong, D. Hajinezhad, and M. Zhao, "Prox-PDA: the proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *Proc. 35th Int. Conf. Mach. Learn*, 2017, pp. 1529–1538. 11

[89] M. Hong, Z. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, 2016. 11

[90] D. Jakovetić, J. M. F. Xavier, and José M. F. Moura, "Convergence rates of distributed nesterov-like gradient methods on random networks.," *IEEE Trans. Signal Process.*, vol. 62, no. 4, pp. 868–882, 2014. 11

[91] F. Saadatniaki, R. Xin, and U. A. Khan, "Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices," *IEEE Trans. Autom. Control*, 2020. 11, 15, 23

[92] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Convergence of asynchronous distributed gradient methods over stochastic networks," *IEEE Trans. on Autom. Control*, vol. 63, no. 2, pp. 434–448, 2018. 11

[93] A. Spiridonoff, A. Olshevsky, and I. C. Paschalidis, "Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions," *J. Mach. Learn. Res.*, vol. 21, no. 58, pp. 1–47, 2020. 11, 137

[94] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Trans. Signal Process.*, vol. 67, no. 19, pp. 4934–4947, Oct 2019. 11

[95] A. Berahas, R. Bollapragada, N. S. Keskar, and E. Wei, "Balancing communication and computation in distributed optimization," *IEEE Trans. on Autom. Control*, 2018. 11

[96] M. Assran and M. Rabbat, "Asynchronous gradient-push," *IEEE Trans. on Autom. Control*, 2020. 11

[97] J. Zhang and K. You, "Asyspa: An exact asynchronous algorithm for convex optimization over digraphs," *IEEE Trans. on Autom. Control*, 2019. 11

[98] B. Swenson, S. Kar, H. V. Poor, and J. M. F. Moura, "Annealing for distributed global optimization," in *Proc. IEEE Conf. Decis. Control*. IEEE, 2019, pp. 3018–3025. 11

[99] A. Daneshmand, G. Scutari, and V. Kungurtsev, "Second-order guarantees of distributed gradient algorithms," *SIAM J. Optim.*, vol. 30, no. 4, pp. 3029–3068, 2020. 11

[100] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—Part I: Agreement at a linear rate," *IEEE Trans. Signal Process.*, vol. 69, pp. 1242–1256, 2021. 11

[101] B. Swenson, R. Murray, H. V. Poor, and S. Kar, "Distributed gradient flow: Nonsmoothness, nonconvexity, and saddle point evasion," *IEEE Trans. Autom. Control*, 2021. 11

[102] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019. 11

[103] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 15

[104] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 15

[105] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. 15

[106] J. M. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar, "Gradient descent on neural networks typically occurs at the edge of stability," *arXiv preprint arXiv:2103.00065*, 2021. 15

[107] H. Wai and A. Scaglione, "Consensus on state and time: Decentralized regression with asynchronous sampling," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2972–2985, 2015. 15

[108] S. Dutta, J. Wang, and G. Joshi, "Slow and stale gradients can win the race," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 3, pp. 1012–1024, 2021. 15

[109] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1383–1400, 2009. 15

[110] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, 2008. 15

[111] Z. Song, L. Shi, S. Pu, and M. Yan, "Compressed gradient tracking for decentralized optimization over general directed networks," *IEEE Trans. Signal Process.*, vol. 70, pp. 1775–1787, 2022. 15

[112] N. Bastianello, R. Carli, L. Schenato, and M. Todescato, "Asynchronous distributed optimization over lossy networks via relaxed admm: Stability and linear convergence," *IEEE Trans. Autom. Control*, vol. 66, no. 6, pp. 2620–2635, 2020. 15

[113] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Math. Program.*, vol. 162, no. 1-2, pp. 83–112, 2017. 17, 20, 21, 30

[114] B. Ying, K. Yuan, and A. H. Sayed, "Dynamic average diffusion with randomized coordinate updates," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 4, pp. 753–767, 2019. 17

[115] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning–Part I: Algorithm development," *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 708–723, 2018. 18, 25, 107

[116] S. Pu, A. Olshevsky, and I. C. Paschalidis, "A sharp estimate on the transient time of distributed stochastic gradient descent," *IEEE Trans. Autom. Control*, 2021. 18, 20, 50, 115, 136, 137

[117] Z. Shen, A. Mokhtari, T. Zhou, P. Zhao, and H. Qian, "Towards more efficient stochastic decentralized learning: Faster convergence and sparse communication," in *International Conference on Machine Learning*, 2018, pp. 4624–4633. 18, 25

[118] Z. Wang and H. Li, "Edge-based stochastic gradient algorithm for distributed optimization," *IEEE Trans. Netw. Sci.*, 2019. 18, 25

[119] H. Hendrikx, F. Bach, and L. Massoulié, "An accelerated decentralized stochastic proximal algorithm for finite sums," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 952–962. 18, 25

[120] B. Li, S. Cen, Y. Chen, and Y. Chi, "Communication-efficient distributed optimization in networks with gradient tracking and variance reduction," *J. Mach. Learn. Res.*, vol. 21, no. 180, pp. 1–51, 2020. 18, 25, 51, 79, 85, 140, 144, 174

[121] B. Ying, K. Yuan, and A. H. Sayed, "Variance-reduced stochastic learning under random reshuffling," *IEEE Trans. Signal Process.*, vol. 68, pp. 1390–1408, 2020. 25

[122] A. Defazio, "A simple practical accelerated method for finite sums," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 676–684. 25

[123] C.-X. Shi and G.-H. Yang, "Augmented Lagrange algorithms for distributed optimization over multi-agent networks via edge-based method," *Automatica*, vol. 94, pp. 55–62, 2018. 25

[124] Q. Lin, Z. Lu, and L. Xiao, "An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization," *SIAM J. Optim.*, vol. 25, no. 4, pp. 2244–2273, 2015. 25

[125] B. Gharesifard and J. Cortés, "Distributed strategies for generating weight-balanced and doubly stochastic digraphs," *European Journal of Control*, vol. 18, no. 6, pp. 539–557, 2012. 26, 55

[126] J. Zhang and K. You, "Decentralized stochastic gradient tracking for empirical risk minimization," *arXiv preprint arXiv:1909.02712*, 2019. 31, 107, 108, 112

[127] D. Williams, *Probability with martingales*, Cambridge university press, 1991. 35, 65, 102

[128] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009. 49, 140

[129] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Math. Program.*, vol. 176, no. 1-2, pp. 497–544, 2019. 50, 107, 168

[130] J. Lei, H. Chen, and H. Fang, "Asymptotic properties of primal-dual algorithm for distributed stochastic optimization over random networks with imperfect communications," *SIAM J. Control Optim.*, vol. 56, no. 3, pp. 2159–2188, 2018. 51

[131] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *Math. Program.*, vol. 180, no. 1, pp. 237–284, 2020. 51

[132] S. Cen, H. Zhang, Y. Chi, W. Chen, and T. Liu, "Convergence of distributed stochastic variance reduced methods without sampling extra data," *IEEE Trans. Signal Process.*, vol. 68, pp. 3976–3989, 2020. 51, 206

[133] L. M. Nguyen, K. Scheinberg, and M. Takac, "Inexact sarah algorithm for stochastic optimization," *Optimization Methods and Software*, pp. 1–22, 2020. 51

[134] D. Zhou, P. Xu, and Q. Gu, "Stochastic nested variance reduction for nonconvex optimization," *J. Mach. Learn. Res.*, 2020. 51, 141, 147

[135] H. Sun, S. Lu, and M. Hong, "Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking," in *Proc. 37th Int. Conf. Mach. Learn.* PMLR, 2020, pp. 9217–9228. 52, 54, 59, 78, 79, 80, 81, 82, 140, 141, 142, 144, 148, 170, 173

[136] D. Zhou and Q. Gu, "Lower bounds for smooth nonconvex finite-sum optimization," in *Proc. 36th Int. Conf. Mach. Learn.* PMLR, 2019, pp. 7574–7583. 54, 55

[137] A. Antoniadis, I. Gijbels, and M. Nikolova, "Penalized likelihood regression for generalized linear models with non-quadratic penalties," *Annals of the Institute of Statistical Mathematics*, vol. 63, no. 3, pp. 585–615, 2011. 59, 116, 117, 147

[138] M. Gurbuzbalaban, A. Ozdaglar, and P. A. Parrilo, "On the convergence rate of incremental aggregated gradient algorithms," *SIAM J. Optim.*, vol. 27, no. 2, pp. 1035–1048, 2017. 78

[139] H. Wai, W. Shi, C. A. Uribe, A. Nedić, and A. Scaglione, "Accelerating incremental gradient optimization with curvature information," *Comput. Optim. Appl.*, pp. 1–34, 2020. 78

[140] A. Mokhtari, M. Gurbuzbalaban, and A. Ribeiro, "Surpassing gradient descent provably: A cyclic incremental method with linear convergence rate," *SIAM J. Optim.*, vol. 28, no. 2, pp. 1420–1447, 2018. 78

[141] R. Xin, U. A. Khan, and S. Kar, "Fast decentralized nonconvex finite-sum optimization with recursive variance reduction," *SIAM J. Optim.*, vol. 32, no. 1, pp. 1–28, 2022. 78, 79, 80, 81, 82, 84, 87, 140, 144, 152, 169, 170, 173, 178, 179

[142] H. Li, Z. Lin, and Y. Fang, "Optimal accelerated variance reduced EXTRA and DIGing for strongly convex and smooth decentralized optimization," *arXiv preprint arXiv:2009.04373*, 2020. 79, 85, 140, 178

[143] Y. Tang, J. Zhang, and N. Li, "Distributed zero-order algorithms for nonconvex multi-agent optimization," *IEEE Trans. Control. Netw. Syst.*, 2020. 79, 85, 108, 114

[144] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Linear convergence of first-and zeroth-order primal-dual algorithms for distributed nonconvex optimization," *IEEE Trans. Autom. Control*, 2021. 79, 85

[145] S. J. Reddi, S. Sra, B. Poczos, and A. J. Smola, "Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1145–1153. 79, 81, 85

[146] L. Zhao, M. Mammadov, and J. Yearwood, "From convex to nonconvex: a loss function analysis for binary classification," in *IEEE International Conference on Data Mining Workshops*, 2010, pp. 1281–1288. 81, 86, 168, 179

[147] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 795–811. 81, 87, 110, 118

[148] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "A primal-dual SGD algorithm for distributed nonconvex optimization," *IEEE/CAA J. Autom. Sin.*, vol. 9, no. 5, pp. 812–833, 2022. 85, 140, 142, 144, 148

[149] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth, "Lower bounds for non-convex stochastic optimization," *Math. Program.*, pp. 1–50, 2022. 106, 141, 169, 173, 177

[150] S. Lu, X. Zhang, H. Sun, and M. Hong, "GNSD: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization," in *2019 IEEE Data Science Workshop*, 2019, pp. 315–321. 107, 108, 112, 140, 144, 148, 173

[151] S. Kar and José M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 674–690, 2011. 108, 115, 136, 137

[152] H. Robbins and D. Siegmund, "A convergence theorem for non negative almost supermartingales and some applications," in *Optimizing methods in statistics*, pp. 233–257. Elsevier, 1971. 108, 133

[153] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. 35th Int. Conf. Mach. Learn.*, 10–15 Jul 2018, pp. 1467–1476. 110

[154] M. B. Nevelson and R. Z. Hasminskii, *Stochastic approximation and recursive estimation*, vol. 47, American Mathematical Soc., 1976. 114, 115, 136

[155] H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Quantized decentralized stochastic learning over directed graphs," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 9324–9333. 140

[156] H. Li and Z. Lin, "Revisiting EXTRA for smooth distributed optimization," *SIAM J. Optim.*, vol. 30, no. 3, pp. 1795–1821, 2020. 140

[157] C. Xi, R. Xin, and U. A. Khan, "ADD-OPT: Accelerated distributed directed optimization," *IEEE Trans. Autom. Control*, vol. 63, no. 5, pp. 1329–1339, 2017. 140

[158] T. Pan, J. Liu, and J. Wang, "D-SPIDER-SFO: A decentralized optimization algorithm with faster convergence rate for nonconvex problems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 1619–1626. 140, 141, 142, 144, 148

[159] K. Rajawat and C. Kumar, "A primal-dual framework for decentralized stochastic optimization," *arXiv preprint arXiv:2012.04402*, 2020. 140

[160] R. Xin, U. A. Khan, and S. Kar, "A fast randomized incremental gradient method for decentralized non-convex optimization," *IEEE Trans. Autom. Control*, 2021. 140, 144

[161] Q. Lü, X. Liao, H. Li, and T. Huang, "A computation-efficient decentralized algorithm for composite constrained optimization," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 774–789, 2020. 140

[162] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen, "A hybrid stochastic optimization framework for stochastic composite nonconvex optimization," *Math. Program.*, 2020. 140, 141, 144, 147

[163] A. Cutkosky and F. Orabona, "Momentum-based variance reduction in non-convex SGD," in *Adv. Neural Inf. Process. Syst.*, 2019, pp. 15236–15245. 140, 141, 144

[164] R. Xin, S. Das, U. A. Khan, and S. Kar, "A stochastic proximal gradient framework for decentralized non-convex composite optimization: Topology-independent sample complexity and communication efficiency," *arXiv preprint arXiv:2110.01594*, 2021. 166

[165] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized sgd with changing topology and local updates," in *Proc. 37th Int. Conf. Mach. Learn.* PMLR, 2020, pp. 5381–5393. 167

[166] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951. 167, 174

[167] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*, SIAM, 2014. 167

[168] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5239–5269, 2019. 168

[169] W. Shi, Q. Ling, G. Wu, and W. Yin, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6013–6023, 2015. 168

[170] S. Alghunaim, K. Yuan, and A. H. Sayed, "A linearly convergent proximal gradient algorithm for decentralized optimization," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 2848–2858, 2019. 168

[171] Y. Sun, G. Scutari, and A. Daneshmand, "Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation," *SIAM J. Optim.*, vol. 32, no. 2, pp. 354–385, 2022. 168, 174

[172] Y. Li, X. Liu, J. Tang, M. Yan, and K. Yuan, "Decentralized composite optimization with compression," *arXiv preprint arXiv:2108.04448*, 2021. 168

[173] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2834–2848, 2018. 168

[174] C. Chen, J. Zhang, L. Shen, P. Zhao, and Z. Luo, "Communication efficient primal-dual algorithm for nonconvex nonsmooth distributed optimization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1594–1602. 168

[175] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 2, pp. 391–405, 2012. 168

[176] Z. Wang, J. Zhang, T. Chang, J. Li, and Z. Luo, "Distributed stochastic consensus optimization with momentum for nonconvex nonsmooth problems," *IEEE Trans. Signal Process.*, 2021. 168, 179, 180

[177] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Math. Program.*, vol. 155, no. 1-2, pp. 267–305, 2016. 169, 174, 175, 177

[178] R. Xin, U. A. Khan, and S. Kar, "A hybrid variance-reduced method for decentralized stochastic non-convex optimization," in *Proc. 38th Int. Conf. Mach. Learn.*, 18–24 July 2021, pp. 11459–11469. 169, 170, 173, 177, 179

[179] Y. Lu and C. De Sa, "Optimal complexity in decentralized training," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 7111–7123. 170, 173, 179

[180] D. Jakovetić, J. Xavier, and J. M. F. Moura, "Fast distributed gradient methods," *IEEE Trans. Autom. Control*, vol. 59, no. 5, pp. 1131–1146, 2014. 174

[181] A. Olshevsky, "Linear time average consensus and distributed optimization on fixed graphs," *SIAM J. Control Optim.*, vol. 55, no. 6, pp. 3990–4014, 2017. 178